LREC 2020 Workshop
Language Resources and Evaluation Conference
11–16 May 2020

9th Workshop on the Representation and Processing of Sign Languages:
Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives

PROCEEDINGS

Eleni Efthimiou, Stavroula-Evita Fotinea,Thomas Hanke,
Julie A. Hochgesang, Jette Kristoffersen, Johanna Mesch (eds.)

Proceedings of the LREC 2020
9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives

Edited by: Eleni Efthimiou, Stavroula-Evita Fotinea,Thomas Hanke, Julie A. Hochgesang, Jette Kristoffersen, and Johanna Mesch

ISBN: 979-10-95546-54-2
EAN: 9791095546542

For more information:
European Language Resources Association (ELRA)
9 rue des Cordelières
75013, Paris
France
http://www.elra.info
Email: lrec@elda.org

© European Language Resources Association (ELRA)

These workshop proceedings are licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

Preface

This collection of papers stems from the 9th Workshop on the Representation and Processing of Sign Languages which was supposed to be held in May 2020 as a satellite to the Language Resources and Evaluation Conference in Marseille (France). While the workshop itself had to be postponed due to the Corona Virus, these proceedings go online as planned as a service to the community.

While there has been occasional attention to sign languages at the main LREC conference, the focus there is on spoken languages in their written and spoken forms. This series of workshops, however, offers a forum for researchers focussing on sign languages, especially on corpus data and corpus technology for sign languages.

This year's hot topic "Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Services" strongly reminds us that our field should progress in a way that maximally benefits the language communities. It is also our obligation to explain where long-term research on major technological challenges may finally contribute to overwhelming barriers the communities are still confronted with in collaborative ways where preferably they as main stakeholders are leading the efforts themselves whenever possible.

The contributions composing this volume are presented in alphabetical order by the first author. For the reader's convenience, an author index is provided as well.

Once again, we would like to thank all members of the program committee who helped us tremendously by reviewing the submissions to the workshop within a very short timeframe!

Finally, we would like to point the reader to the proceedings of the previous workshops that form important resources in a growing field of research. They are all available online from
http://www.sign-lang.uni-hamburg.de/lrec/
The site now also offers an author index across all workshops as well as DOIs for all workshop papers and posters. If you need BibTex data for all workshops, the site now has them per paper, per workshop, per author or all in one. Happy browsing!

The Editors

Organizers:

Eleni Efthimiou, Institute for Language and Speech Processing, Athens, Greece
Stavroula-Evita Fotinea, Institute for Language and Speech Processing, Athens, Greece
Thomas Hanke, Institute of German Sign Language, University of Hamburg, Germany
Julie Hochgesang, Gallaudet University, Washington, USA
Jette Kristoffersen, Centre for Sign Language, Copenhagen, Denmark
Johanna Mesch, Stockholm University, Sweden

Program Committee:

Annelies Braffort, LIMSI/CNRS, Orsay, France
Onno Crasborn, Radboud University, Nijmegen, The Netherlands
Sarah Ebling, University of Zurich, Zurich, Switzerland
Eleni Efthimiou, Institute for Language and Speech Processing, Athens, Greece
Michael Filhol, LIMSI/CNRS, Orsay, France
Stavroula-Evita Fotinea, Institute for Language and Speech Processing, Athens, Greece
Thomas Hanke, University of Hamburg, Hamburg, Germany
Julie A. Hochgesang, Gallaudet University, Washington, USA
Tommi Jantunen, University of Jyväskylä, Jyväskylä, Finland
Trevor Johnston, Macquarie University, Sydney, Australia
Reiner Konrad, University of Hamburg, Hamburg, Germany
Jette Kristoffersen, Centre for Sign Language, Copenhagen, Denmark
John McDonald, DePaul University, Chicago, USA
Johanna Mesch, Stockholm University, Stockholm, Sweden
Carol Neidle, Boston University, Boston, USA
Marc Schulder, University of Hamburg, Hamburg, Germany
Rosalee Wolfe, DePaul University, Chicago, USA

Table of Contents

Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 1–6
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

Back and Forth between Theory and Application: Shared Phonological Coding Between ASL Signbank and ASL-LEX

Amelia Becker[1], Donovan H. Catt[2], Julie A. Hochgesang[2]
Georgetown University[1], Gallaudet University[2]
Washington, DC, USA
ab3082@georgetown.edu; {donovan.catt; julie.hochgesang}@gallaudet.edu

Abstract

The development of signed language lexical databases, digital organizations that describe different phonological features of and attempt to establish relationships between signs has resulted in a renewed interest in the phonological descriptions used to uniquely identify and organize the lexicons of respective sign languages (van der Kooij, 2002; Fenlon et al., 2016; Brentari et al., 2018). Throughout the mutually shared coding process involved in organizing two lexical databases, ASL Signbank (Hochgesang, Crasborn and Lillo-Martin, 2020) and ASL-LEX (Caselli et al., 2016), issues have arisen that require revisiting how phonological features and categories are to be applied and even decided upon, and which would adequately distinguish lexical contrast for respective sign languages. The paper concludes by exploring the inverse of the theory-to-database relationship. Examples are given of theoretical implications and research questions that arise from consequences of language resource building. These are presented as evidence not only that theory impacts organization of databases but that the process of database creation can also inform our theories.

Keywords: lexical database, phonological coding, signed languages, ASL, experiences in building sign language corpora, annotation and visualization tools

1. Introduction

The development of signed language lexical databases, digital organizations that describe different phonological features of and attempt to establish relationships between signs has resulted in a renewed interest in the phonological descriptions used to uniquely identify and organize the lexicons of respective sign languages (van der Kooij, 2002; Fenlon et al., 2016; Brentari et al., 2018). Throughout the mutually shared coding process involved in organizing two lexical databases, ASL Signbank (Hochgesang, Crasborn and Lillo-Martin, 2020) and ASL-LEX (Caselli et al., 2016), issues have arisen that require revisiting how phonological features and categories are to be applied and even decided upon, and which would adequately distinguish lexical contrast for respective sign languages.

One way in which issues arise in applying phonological descriptions is how we should go about making explicit categories that subsume several "descriptors" or features: movement paths include *arc/curved, straight, circular,* and *other*. This 'other' category includes multiple path types, such as repeated straight path movements, back and forth (bidirectional) movement, or a circular or curved path followed by a straight path movement combination. Furthermore, determining the nature of the type of multiple paths for a given sign is necessary in distinguishing this from holistic views of path movements: e.g. the difference between the ASL signs for DUTYtap[1] and COMMUTE[2] (both are currently identified as [BackAndForth], but we recognize that DUTYtap actually repeats [Straight] path movements, while COMMUTE is a true example of [BackAndForth]).

For a lexical database this kind of information needs to be explicit in order for the function/purpose of distinguishing signs, as well as for comparing how signs are similar. For the ASL Signbank, where the phonological coding of each lexical entry is carried out in collaboration with ASL-LEX lexical database, the fact that there are numerous false homonyms is one such issue that this project aims to address.

2. Background

2.1 Lexical databases and phonological neighborhoods

Lexical databases are developed for myriad purposes, including lexicography and as resources for investigating the phonological structure of a given language. To the end of the latter purpose, which is the focus of this paper, lexical databases have traditionally been organized in such a way as to visualize phonological neighborhoods, or to show how lexical units are similar (or different) according to phonological properties. Phonological neighbors have traditionally been recognized as having a single phoneme difference between two words/signs (Marian and Blumenfeld, 2006). Both ASL Signbank and ASL-LEX currently utilize an abbreviated version of the Prosodic Model (Brentari, 1998), henceforth abbreviated as PM, a feature-based phonological descriptive system, and in turn recognize that this contributes to (and even affects) the overall organization and how lexical relationships are indicated.

As Hochgesang (2014: 490) explains, "systems of measurement should be thoroughly and consistently vetted before they are adopted for widespread use". Because both the ASL Signbank and ASL-LEX databases are still being expanded, the development process is crucial for working out any apparent issues that could undermine the intended benefits of each resource. Caselli et al. (2016: 790) assert that the phonological coding scheme as applied "has substantial discriminatory power;" they indicated that "52% of signs were uniquely identified, and 32% shared a phonological transcription with fewer than three other signs." Signs are identified as related (shared phonological properties) in ASL-LEX in three ways: those that have the same major location, selected fingers, flexion, and movement (parameter-based neighborhood density), those that share four of five phonological features, which adds in consideration of a sign's minor location along with the previously mentioned properties (maximal neighborhood density); and those that share at least one phonological

[1] Sign ID numbers that link to the ASL Signbank will be given for examples. To find the linked sign, registered users can visit https://aslsignbank.haskins.yale.edu/dictionary/gloss/# replace # with the ID number provided in the footnote. DUTYtap is 1283.

[2] ASL Signbank ID 1109

feature (minimal neighborhood density) (Caselli et al., 2016: 792).

As for the ASL Signbank, shared phonological descriptions appears to be a much more pervasive issue (e.g. entering in [Hand], [imrp], [1 (fully open)], [Straight] path movement within the sign search function leads to a result of 36 signs sharing a phonological transcription, many of which would be considered "false homonyms"). Applying a feature-based system to lexical databases might seem counterintuitive if phonological relationships are determined based on a phonemic descriptive approach, but as Corina (1990: 27) explains, "[i]n describing distinctive feature systems, one attempts to characterize the underlying perceptual and/or gestural components of phonemes in a [sign] language." For example, "a handshape representation consists of features for finger(s) involved in articulating the handshape and features describing the configuration [or flexion] of these fingers" (Corina, 1990: 28).

2.2 ASL-LEX

ASL-LEX[3] (Caselli et al., 2016) is a publicly-available database which includes subjective frequency and iconicity judgments as well as phonological information for 1,000 ASL signs (and more to come in version 2.0).

2.3 ASL Signbank

The ASL Signbank[4], further described in Hochgesang, Crasborn and Lillo-Martin (2018), is an online database that organizes ID glosses for ASL annotation. It is built off the NGT Signbank, which in turn is based on the Auslan Signbank software (Cassidy et al., 2018). At the time of writing, there are over 3300 entries. ASL Signbank is a language resource that can be directly linked to ELAN (Crasborn et al., 2016).

2.4 Phonological coding used by ASL Signbank

Our collaboration between ASL Signbank and ASL-LEX involves sharing ID glosses, so that signs that are common across the databases can be easily accessed, as well as phonological information, with the goal that signs have consistent coding, whenever applicable, across the two databases (discrepancies in coding between the two databases helped to illuminate some of the issues discussed in section 3 below). The phonological coding scheme used for both the ASL Signbank and ASL-LEX are based on the PM which is essentially a compilation of the collective analyses that have been carried out on ASL phonology over the past few decades (Brentari, 1998). In compound signs, codes refer to the properties of the initial free morpheme (or component derived from what was originally the initial free morpheme). Note that the phonological coding in ASL Signbank does not provide a complete phonological description of signs; there are contrastive elements that are not included for entries, e.g. direction of movement. This leads to situations in which signs that are distinct in form and meaning are identically coded for phonology in ASL Signbank (e.g. ACT[5] and AGGRESSIVE[6], which differ only in direction of movement, a characteristic which is not coded here). In this subsection, we describe some of the fields in which we apply the shared phonological coding scheme for both the ASL Signbank and ASL-LEX in the Phonology section of the ASL Signbank. Not all phonological aspects included in ASL Signbank will be described here (e.g. not describing *weak drop/prop* since this is still in development) – we'll describe handedness, major location and dominant hand – selected fingers and flexion.

2.4.1 Handedness

ASL signs can be one or two-handed. When two-handed, they tend to conform to constraints referred to as the Symmetry and Dominance Conditions (Battison, 1978). According to the Symmetry Condition, signs for which both hands move must have the same or mirror image location and orientation, same handshape, and same (simultaneous or alternating) movement specifications. This type of sign is listed in ASL Signbank as [SymmetricalOrAlternating] (e.g. ACCEPT[7] has symmetrical specifications and BICYCLE[8] has alternating movement). Signs in which only one hand moves are referred to as [asymmetrical]. When the two handshapes are the same in an asymmetrical sign, these signs are coded in ASL Signbank as [AsymmetricalSameHandshape] (e.g. BELIEVEb[9]). According to the Dominance Condition, when a two-handed sign has different specifications for the two handshapes, the sign must be asymmetrical (that is, only one hand can move), and the stationary hand is restricted to one of seven unmarked handshapes, coded in the Nondominant handshape field as 1, 5, A, B, C, O, and S. These signs are coded in ASL Signbank as [AsymmetricalDifferent Handshape] (e.g. COUNT[10]). Finally, two-handed signs may be coded in ASL Signbank as [Other] when they violate either the Symmetry or Dominance Condition. Signs that violate the Symmetry Condition are those for which both hands move but have different handshapes (e.g. SIM-COM[11]). Signs that violate the Dominance Condition are those for which the stationary hand has a handshape other than the seven unmarked handshapes (e.g. CHERRY[12]). The possible values for handedness in ASL Signbank are listed below:

AsymmetricalDifferentHandshape
AsymmetricalSameHandshape
OneHanded
Other (violates sym/dom conditions)
SymmetricalOrAlternating

2.4.2 Location – Major

Each sign is specified for only one major location. The possible locations are listed below, along with examples. Note that there does not need to be contact between the hand and location, either in phonological specification or in actual production. For each pair of examples, the first makes contact with the major location and the second does not (excluding neutral, for which there can never be contact with the body). The possible values are listed here:

arm (including wrist): e.g. TRASH[13]
body (signer's torso): e.g. FANCY[14]

[3] http://asl-lex.org/
[4] https://aslsignbank.haskins.yale.edu/
[5] ASLSignbank Sign ID 5
[6] ASLSignbank Sign ID 2379
[7] ASLSignbank Sign ID 2045
[8] ASLSignbank Sign ID 379
[9] ASLSignbank Sign ID 576
[10] ASLSignbank Sign ID 609
[11] ASLSignbank Sign ID 3289
[12] ASLSignbank Sign ID 1982
[13] ASLSignbank Sign ID 2107
[14] ASLSignbank Sign ID 1382

hand: e.g. BEACHwig[15], BASIC[16]
head (including face): ALASKA[17]
neutral (signing space in front of the signer's body): e.g. INSULT[18]
other NA

2.4.3 Dominant hand – Selected Fingers and Flexion

ASL signs adhere to "the Finger Position Constraint" (Mandel, 1981) which limits the number of categories a handshape can specify for finger configurations to two. One group of fingers – called the selected fingers – can be specified for any configuration possible in ASL. The other group – the non-selected fingers – must be either fully extended or fully flexed/closed. This means, for example, that a handshape in which some fingers are specified as fully extended, some as partially extended, and some as fully flexed is impossible in ASL. Since Mandel (1981), various models have formalized this constraint in slightly different ways; all capture the notion that signs specify one category of phonologically salient fingers. ASL Signbank follows ASL-LEX's and PM's criteria for coding selected fingers. In signs with a handshape change or handshape-internal movement, the fingers that move are selected (e.g. index in QM[19]). For signs without a handshape change or handshape-internal movement, if one set of fingers is partially flexed or partially extended (e.g. index in NEED[20]), these fingers are considered selected and the set of fully flexed or fully extended fingers are considered non-selected. If neither of these criteria can be applied to distinguish between selected and non-selected fingers, the decision is made based on which fingers "appear foregrounded" (Caselli et al., 2016). For example, in the sign ALONE[21] there is no handshape change or internal movement, one category of fingers (index) is fully extended, and the other category (middle, ring, and pinky) is fully flexed. Since neither group is partially extended/flexed, stacked, or crossed, applying these criteria does not differentiate selected from non-selected fingers in ALONE. However, the index finger appears foregrounded and is therefore coded as the selected finger.

The PM model states that "in the majority of cases the thumb behaves like the other selected fingers...yet in some signs it operates as a semi-independent articulator" (Brentari, 1998, p. 113). In ASL-LEX and ASL Signbank, the thumb is coded as selected only when it is the only selected finger. For example, in MOON[22], both index and thumb are partially extended and middle, ring, and pinky are fully flexed. In this case, only the index is coded as the selected finger. In the sign TEXT-PAGER[23], on the other hand, since the thumb is the only moving/salient finger while the others are fully flexed and non-moving, the thumb is coded as the selected finger. In asymmetrical two-handed signs, selected fingers are coded only for the dominant hand (e.g. middle is selected for ADVANTAGE[24]). The full word [thumb] labels the thumb as the selected finger. The codes for the remaining fingers are each one letter: i = index, m = middle, r = ring, and p = pinky. All possible combinations of the four fingers, including each finger individually, are possible in this field with the exception of ir (index and ring), mp (middle and pinky), and rp (ring and pinky), which are unattested in ASL.

Following ASL-LEX and PM, flexion codes in ASL Signbank are categorical. That is, rather than providing a phonetic description of the flexion of individual joints, flexion codes describe nine categories of hand configurations that arise from combinations of flexion values of selected finger joints and configuration of the thumb in relation to the selected fingers. Selected finger joints may be "flat", "bent", or "curved." In "flat" configurations, selected fingers are flexed at the metacarpal joints only. In "bent" configurations, the distal and proximal joints are flexed. "Curved" configurations are those in which the selected finger joints are partially flexed. The thumb can be either "closed", in which case it contacts the fingers, or "open", in which case it does not. These finger and thumb configurations combine to produce seven contrastive categories. Two additional joint configurations – crossing and stacking – provide the last two possible values in the Flexion field.

When flexion changes due to handshape change or handshape-internal movement, only the initial state is coded. For asymmetrical two-handed signs, the values given in this field reflect the dominant hand configuration only. Below, each contrastive category resulting from the finger and thumb configurations just presented, is described, and an example is given. The first seven are coded in ASL Signbank by a numerical label, and the last two are simply named [Crossed] and [Stacked]. The possible values are listed below with examples:

1: fully open – finger joints fully extended and thumb unopposed, not contacting fingers (e.g. ABHOR[25])
2: bent or closed – (e.g. BATTERY[26])
3: flat open – metacarpal joints flexed, thumb not contacting fingers (e.g. GROWN-UP[27])
4: flat closed – metacarpal joints fully flexed, thumb contacting selected or non-selected fingers (e.g. BUY[28])
5: curved open – finger joints partially flexed, thumb not contacting fingers (e.g. HOT[29])
6: curved closed – finger joints partially flexed, thumb contacting fingers (e.g. EIGHT[30])
7: fully closed – finger joints fully flexed, thumb may or may not be contacting fingers (e.g. SHOES[31])
Crossed (e.g. DONUTneut[32]) – selected fingers crossed over one another (e.g. ROPE[33])

[15] ASLSignbank Sign ID 583
[16] ASLSignbank Sign ID 2092
[17] ASLSignbank Sign ID 707
[18] ASLSignbank Sign ID 2970
[19] ASLSignbank Sign ID 529
[20] ASLSignbank Sign ID 194
[21] ASLSignbank Sign ID 2065
[22] ASLSignbank Sign ID 1594
[23] ASLSignbank Sign ID 1519
[24] ASLSignbank Sign ID 2052
[25] ASLSignbank Sign ID 2040
[26] ASLSignbank Sign ID 345
[27] ASLSignbank Sign ID 2442
[28] ASLSignbank Sign ID 424
[29] ASLSignbank Sign ID 485
[30] ASLSignbank Sign ID 635
[31] ASLSignbank Sign ID 393
[32] ASLSignbank Sign ID 2419
[33] ASLSignbank Sign ID 663

Stacked (e.g. WORSE[34]) – different flexion value for each selected finger (e.g. ALLOWp[35])

3. Some specific issues with current phonological coding system

Through assessment of the application of phonological properties to both databases several discrepancies arose. As examples, we will focus here on two: one pertaining to finger flexion and one related to distinguishing path movement types. An example of when a flexion coding discrepancy becomes apparent is determining whether a sign should be considered [3] "flat open" versus [4] "flat closed". The other coding discrepancy discussed here relates to how movement types are distinguished, such as signs with repeated straight path movements as opposed to a back and forth movement, or signs with a repeated arc path movement as opposed to a circular one.

3.1 Distinguishing flexions

When considering the feature specifications for the thumb and index finger (note, this is currently not a possible selected finger combination in ASL Signbank, although it is actually a very common occurrence in signs, because the thumb is only considered selected when it is the only selected finger in our coding scheme), the options [ti] "flat closed", [ti] "curved closed" do not serve a distinctive function (likewise for any thumb and other single finger selection). For both of these the "secondary finger selections" can be either extended or closed, which would seem to result in four possible finger selection-flexion combinations, or handshapes; in reality, there are two: 'F' handshape or a so-called 'baby-O' handshape (with thumb and index contacting). What seems to be the issue for signs within the [ti] and either flexion specification (123 in total) is that the flexion feature is not contrastive (there are 198 signs in the ASL Signbank with the thumb and any other single finger selection and either flexion specification). One possible explanation for this could be that for signs that incorporate an unmarked handshape, flexion serves a more distinctive role, while in signs with marked handshapes, flexion is less pronounced. Or, when signs select for primary and secondary fingers, the primary finger contrast simply lies with closed or open features, while in signs that select for primary fingers only (no secondary selection), the features flat and curved provide additional needed contrast. Interestingly, when describing each flexion feature in the PM, each sign referenced involves all of the fingers ([timrp]), with the flexion categories examined here represented through KNOW-NOTHINGf[36] (curved closed) and KISS-MOUTHstr[37] (flat closed) (Brentari, 1998: 108).

3.2 Path movement discrepancies

The other phonological coding issue examined here relates to path movement discrepancies, and how path movement codes are (inconsistently) applied. The PM identifies movement types (*straight, arc,* and *circle*) and movement sequences (Brentari, 1998: 132). This latter category is where many issues arise in determining whether a sign's path movement should be considered to be repeated straight paths or a bidirectional path. For example, DUTYtap should be a repeated [Straight] path movement rather than bidirectional (i.e., [BackAndForth]; e.g. COMMUTE). Also, for signs like SEARCH[38], should this be a repeated upward [curved] path movement rather than circular (e.g. YEARs[39])? Furthermore, the PM does not discuss the movement sequence type described for signs such as SEARCH, or those that have repeated arc path movements.

Issues with the path movement coding scheme seem in part due to the application of features based on either a perceptual approach (more "global", or path movements taken altogether) versus how they are characterized and sometimes referenced with sign examples in the PM (Brentari 1998). As our coding scheme now stands, identifying DUTYtap as [Straight] seems "off", but this could easily be resolved by adding a [Repeat] feature of sorts. Explicating the distinction between movement sequences is necessary in order to uniquely characterize signs, and, in turn, will affect sign relation results, which is a primary focus of lexical databases.

3.3. General discussion of issues

The implications we can take away from these issues in using specific theories when applying to lexical database organization just outlined in 3.1 and 3.2 are that we need to consider pursuing a recursive/symbiotic relationship between theory and database-building. Issues that arise in database building can inform revision of theory – i.e. when the data contradict the theory. Database building can also support/confirm predictions made by theory.

Assessing both the ASL Signbank and ASL-LEX is particularly significant and necessary for further research based on lexical databases because the phonological coding system applied is, in some respects, a shorthand version of the PM. Even through comprehensive application of the PM, the issue of determining phonological distinctions is still unresolved, and so additional, thorough examination of data (evidence of contrastive units) would be beneficial for both theory and application (Eccarius & Brentari, 2008; Fenlon, Cormier, & Brentari, 2017). This has been acknowledged in other studies on sign language lexicons, such as BSL. In a study outlining the phonological structure of BSL through a usage-based lexical database, Fenlon et al. (2016: 39), concluded that "[o]ther issues to explore in more detail involve searching for evidence of lexical contrast". Furthermore, as explained by Fenlon, Cormier, and Brentari (2017), "[m]ore evidence is needed about lexical contrast in ASL and BSL before claims about particular contrastive units can be confirmed," and so the discussion of issues and discrepancies in this study is an initial step in that direction.

4. Implications for Theory

The examples we have seen so far might be considered conflicts or problems arising from application of a

[34] ASLSignbank Sign ID 1462
[35] ASLSignbank Sign ID 12
[36] ASLSignbank Sign ID 1693
[37] ASLSignbank Sign ID 165
[38] ASLSignbank Sign ID 537
[39] ASLSignbank Sign ID 656

particular theory or coding scheme. However, theory and resource building can, and ideally do, exist in a recursive, symbiotic relationship. Theory provides a foundation for the coding that makes a resource searchable and quantifiable; the act of coding then serves as a test of a theory's predictions, informing revisions where issues arise and confirming those predictions where application is successful. Importantly, whether a coding scheme is theoretically grounded or simply anticipates how a user may want to search a database, cases where its application is less than straightforward often lead to interesting research questions that can be addressed empirically. The remainder of this paper discusses a few examples of this complementary relationship between theory and database coding. The purpose of this section is not necessarily to provide specific solutions to the issues raised in the foregoing discussion but, rather, to present examples of ways in which insights from database building can reveal paths for theoretical research.

One area of ASL Signbank that poses a challenge is categorization of the handshape for the nondominant hand in asymmetrical two-handed signs (those in which the nondominant hand remains stationary). When the two handshapes differ, the coding scheme inherited from ASL-LEX, following Battison's (1978) typology of two-handed signs, restricts the nondominant hand handshape to one of seven possibilities: 1 (e.g. AVOIDix[40]), 5 (e.g. POLICY[41]), A (e.g. TECH[42]), B (e.g. DOLLAR[43]), C (e.g. GET-IN[44]), O (e.g. QUIT[45]), or S (e.g. APPOINTMENT[46]). When the nondominant hand of an asymmetrical two-handed sign cannot be categorized as one of these seven options, the sign is considered to violate Battison's Dominance Condition and is labelled as "Other" in ASL Signbank. Applying this catchall category will allow us, once enough data are collected, to ask questions about what leads to violations of the Dominance Condition and in what ways it can be violated. For example, in the ASL Signbank production of CHAIR[47], the dominant hand has index and middle finger partially extended (curved), while the nondominant handshape has index and middle fully extended.

In another example of a Dominance Condition violation, HELICOPTERthree[48], the nondominant hand handshape has thumb, index, and ring finger fully extended (), while the dominant hand has a 5 handshape .

Both of these examples are still closely related to depictive origins. Furthermore, in CHAIR the two handshapes differ in flexion only and, in HELICOPTERthree, in selected finger combination only. Thus we might ask whether a sign could violate the Dominance Condition by two handshapes that differ in both flexion and selected finger combination and/or whether violations of the Dominance Condition are always in signs closely resembling depictive origin. The potential to understand iconicity and phonology as opposing forces shaping signed languages is a central question in the field which can be probed by cases like these[49].

A second question arises from another issue that is highlighted by categorization of the nondominant hand handshape in ASL Signbank: namely, how to treat signs in which a site on the opposite arm or forearm serves as the location for the dominant hand, but the nondominant hand may or may not be specified for a particular handshape. It is unclear whether it is appropriate to categorize these signs within the handedness typology currently available in ASL Signbank and/or to specify a handshape for the nondominant hand. For example, the sign CRACKER[50] is produced in ASL Signbank with an A handshape on both hands, but it is the elbow location that seems phonologically relevant rather than the nondominant hand handshape.

It can also be difficult to determine the boundary between arm as location, as appears to be the case in CRACKER, and nondominant hand as weak hand if the dominant hand articulates on the wrist or near the base of the hand (e.g. TIME[51]).

Collection and coding of more entries of these types will allow us to address these questions regarding handedness, and it is the application of a set of categories which brings these ambiguous cases into relief.

Of course, not all methods of database organization are based in linguistic theory; some organization applied for purposes of searchability do not necessarily reflect assertions about the grammar but rather how researchers anticipate users may want to search a database. Nevertheless, these methods of organization can lead to questions about lexical categories.

For example, the "Relations to Other Signs" field in ASL Signbank marks connections between signs that users may be interested in but which may or may not reflect relationships in the grammar. Some relations have established definitions, e.g. "Synonym" and "Variant." The catchall category "See Also" links signs in ways that may not yet be clearly understood, which can lead to interesting research questions about lexical organization that can be tested empirically. One such relationship is initialized signs sharing iconic motivation. For example, one paradigm of semantically related signs all share an iconically motivated symmetrical arc movement on each hand produced in neutral space. These signs are differentiated in form only by their handshapes, which correspond to those of the ASL fingerspelling system, reflecting the first letter of a written English translation of each sign's meaning. The signs in ASL Signbank belonging to this group are AGENCY[52],

[40] ASLSignbank Sign ID 2411
[41] ASLSignbank Sign ID 913
[42] ASLSignbank Sign ID 1183
[43] ASLSignbank Sign ID 1268
[44] ASLSignbank Sign ID 778
[45] ASLSignbank Sign ID 786
[46] ASLSignbank Sign ID 2074
[47] ASLSignbank Sign ID 378
[48] ASLSignbank Sign ID 481
[49] We are thankful to an anonymous reviewer for raising this point.
[50] ASLSignbank Sign ID 647
[51] ASLSignbank Sign ID 1233
[52] ASLSignbank Sign ID 2055

CLASS[53], GROUPg[54], FAMILY[55], LEAGUE[56], ORGANIZATION[57], SOCIETY[58], and TEAM[59] (as well as the ASL sign for "union", not currently listed in ASL Signbank) as shown in Figure 1.

Figure 1: Images of signs that share iconically motivated movement (ASL Signbank, 2020)

Although these relationships are marked only for the sake of searchability, we might ask whether their relationship is metalinguistic and diachronic only or whether it holds some synchronic reality in the lexicon. (For discussion of lexical categories based on iconic motivation, see Occhino 2017, and for evidence of synchronic relationship between ASL signs and written English words, see Morford et al. 2011). This question could be explored through experimental means such as lexical priming tasks. Again, it is the imposition of structure on a database that raises this question, and thus an example of how resource building can lead to theoretical linguistic investigation.

5. Bibliographical References

Battison, R. (1978). *Lexical borrowing in American Sign Language*. Silver Spring, MD: Linstock Press.

Brentari, D. (1998). A prosodic model of sign language phonology. Cambridge, MA: MIT Press.

Brentari, D., Fenlon, J., & Cormier, K. (2018). Sign language phonology. *Oxford Research Encyclopedias*. DOI: 10.1093/acrefore/9780199384655.013.117

Caselli, N.K., Sehyr, Z.S., Cohen-Goldberg, A.M., & Emmorey, K. (2016). ASL-LEX: A lexical database of American Sign Language. *Behavior Research Methods.*

Cassidy, S., Crasborn, O., Nieminen, H., Stoop, W., Hulsbosch, M., Even, S., Komen, E., & Johnston, T. (2018). Signbank: Software to support web based dictionaries of sign language. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S.Piperidis, ... T. Tokunaga (Eds.), *LREC 2018, Eleventh International Conference on Language Resources and Evaluation* (pp. 2359-2364). European Language Resources Association.

Corina, D. (1990). Handshape assimilations in hierarchical phonological representation. In C. Lucas (Ed.), *Sign Language Research: Theoretical Issues*, 27-49. Washington, DC: Gallaudet University Press.

Crasborn, O., Bank, R., Zwisterlood, I., van der Kooij, E., Schüller, A., Ormel, E., Nauta, E., van Zuilen, M., van Winsum, F., & Ros, J. (2016). Linking lexical and corpus data for sign languages: NGT Signbank and the Corpus NGT. In E. Efthimiou et al. (Eds.), Proceedings of the 7th Workshop on the Representation and Processing of Sign Langauges: Corpus Mining (pp. 41-46). Portorož, Slovenia: ELRA.

Eccarius, P. & Brentari, D. (2008). Handshape coding made easier: A theoretically based notation for phonological transcription. *Sign Language & Linguistics*, 11(1), 69-101.

Fenlon, J., Cormier, K., Adam, R., & Woll, B. (2016). Issues in determining the phonological structure of sign languages in usage-based lexicons: The case of BSL. Signbank. [under revision]

Fenlon, J., Cormier, K., & Brentari, D. (2017). The phonology of sign languages. In S. J. Hannahs, & A. R. K. Bosch (Eds.), *The Routledge Handbook of Phonological Theory* (pp. 453-475). Routledge.

Hochgesang, J.A. (2014). Using design principles to consider representation of the hand in some notation systems. *Sign Language Studies*, 14(4), 488-542. Washington, DC: Gallaudet University Press.

Hochgesang, J.A., Crasborn, O., & Lillo-Martin, D. (2018). Building the ASL Signbank: Lemmatization principles for ASL. In M. Bono, E. Efthimiou, S.E. Fotinea, T. Hanke, J. Hochgesang, J. Kristoffersen, J. Mesch, & Y. Osugi (eds.) 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community (pp. 69-74). Paris: ELRA.

Mandel, M. (1981). *Phonotactics and Morphophonology in American Sign Language*. Retrieved from https://escholarship.org/uc/item/90v1j5kx

Marian, V. & Blumenfeld, H.K. (2006). Phonological neighborhood density guides: Lexical access in native and non-native language production. *Journal of Social and Ecological Boundaries*, 2(1), 3-35.

Morford, J. P., Wilkinson, E., Villwock, A., Piñar, P., & Kroll, J. F. (2011). When deaf signers read English: Do written words activate their sign translations?. *Cognition*, 118(2), 286-292. https://doi.org/10.1016/j.cognition.2010.11.006

Occhino, C. (2017). An introduction to embodied cognitive phonology: claw-5 hand-shape distribution in ASL and Libras. *Complutense Journal of English Studies*, 25, 69.

Van der Kooij, E. (2002). Phonological categories in Sign Language of the Netherlands: The role of phonetic implementation and iconicity. The Netherlands: LOT.

6. Language Resource References

Hochgesang, J.A., Crasborn, O., & Lillo-Martin, D. (2020). *ASL Signbank*. New Haven, CT: Haskins Lab, Yale University. https://aslsignbank.haskins.yale.edu/

[53] ASLSignbank Sign ID 585
[54] ASLSignbank Sign ID 643
[55] ASLSignbank Sign ID 1379
[56] ASLSignbank Sign ID 1684
[57] ASLSignbank Sign ID 1529
[58] ASLSignbank Sign ID 1092
[59] ASLSignbank Sign ID 1178

Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 7–12
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

Improving and Extending Continuous Sign Language Recognition: Taking Iconicity and Spatial Language into account

Valentin Belissen[1], Michèle Gouiffès[2], Annelies Braffort[3]
[1, 2, 3]LIMSI–CNRS, [1, 2]Université Paris Sud, [1, 2]Université Paris Saclay
[1, 2, 3]LIMSI, Campus universitaire 507, Rue du Belvedère, 91405 Orsay - France
{valentin.belissen, michele.gouiffes, annelies.braffort}@limsi.fr

Abstract

In a lot of recent research, attention has been drawn to recognizing sequences of lexical signs in continuous Sign Language corpora, often artificial. However, as SLs are structured through the use of space and iconicity, focusing on lexicon only prevents the field of Continuous Sign Language Recognition (CSLR) from extending to Sign Language Understanding and Translation. In this article, we propose a new formulation of the CSLR problem and discuss the possibility of recognizing higher-level linguistic structures in SL videos, like classifier constructions. These structures show much more variability than lexical signs, and are fundamentally different than them in the sense that form and meaning can not be disentangled. Building on the recently published French Sign Language corpus Dicta-Sign–LSF–v2, we also discuss the performance and relevance of a simple recurrent neural network trained to recognize illustrative structures.
Keywords: Continuous Sign Language Recognition, Iconicity, Annotation

1. Introduction

In a series of recent papers focused on Continuous Sign Language Recognition (CSLR), the performance seems to be getting better and better and one may wonder what remains to be done before Sign Language Translation (SLT) can be envisioned. However, we want to show that the current trend in CSLR has inherent limitations, which prevent it from extending to SL understanding, *a fortiori* to SLT. Besides, extending on previous work, we try to highlight the value of recognizing illustrative elements in SL for the task of CSLR, even though few corpora are available for this research. This paper will also discuss the related challenges and the difficulties in evaluating the recognition results.

In Section 2, we start with a summary on the current CSLR paradigm, including linguistic assumptions, prevalent corpora and performance metrics. Then, the formulation we propose for a broader CSLR is presented in Section 3, with a discussion on some important linguistic properties of SLs, interesting corpora and a formal description. Building on this more general acceptation of CSLR, we end in Section 4 with an analysis and discussion of the results in the automatic recognition of illustrative structures within the recently published corpus Dicta-Sign–LSF–v2.

2. The current paradigm: Continuous *Lexical Sign* Recognition

In this section, we formalize the current CSLR paradigm. It focuses on the recognition of specific elements called *lexical signs*, therefore we choose to call it Continuous *Lexical Sign* Recognition (C*LS*R). Three main corpora are used, with a similar annotation scheme and types of discourse with limited variability, that do not allow for an easy generalization.

2.1. Lexical signs

In this paper, we refer to lexical signs following the definition of Johnston and Schembri (2007): *"fully-lexical signs are highly conventionalised signs in both form and meaning in the sense that both are relatively stable or consistent across contexts. Fully-lexical signs can easily be listed in a dictionary".*

Whereas it is fairly safe to draw a parallel between common words and lexical signs, it only goes so far. Indeed, it is important to note that the grammar of SL is realized through the use of very different and more complex structures like described in Section 3.1.2.

2.2. Prevalent corpora

Since learning models are determined by the data they are trained with, it is necessary to discuss the nature of prevalent datasets in the field of CLSR. Three corpora stand out, and have been extensively used for training CLSR models:

- Signum (Von Agris and Kraiss, 2007) is a German Sign Language (DGS) dataset, with 5 hours of RGB video from 25 signers. Pre-defined sentences are elicited, with 465 lexical signs.
- RWTH Phoenix Weather 2014 (Forster et al., 2014; Koller et al., 2017) is made from 11 hours of live DGS interpretation of weather forecast on German TV.
- CSL-25k (Huang et al., 2018) results from the elicitation of 100 pre-defined sentences from 50 signers, with 178 annotated lexical signs, in Chinese Sign Language (CSL).

Table 1 presents a few random elicitation or transcribed sentences from these three corpora.

2.3. Type of discourse – elicitation material

Since they consist in elicited pre-defined sentences, it is important to realize that Signum and CSL-25k are, to some extent, artificial and with limited generalizability. In the CSL-25k corpus, the elicited sentences follow a simple syntactic structure, whereas the sentences in Signum show a

Signum	- Excuse me, could you help me? - The teacher will guide a tour through the church on Wednesday.
Phoenix	- The week starts unpredictable and cooler. - On Sunday spreads partly heavy shower and thunderstorm.
CSL-25k	- My dad is a businessman. - The cup is orange.

Table 1: Random elicited or transcribed sentence examples from three common SLR corpora. Signum and CSL-25k are elicited from pre-written sentences, while Phoenix is a live interpretation from weather forecast on German TV.

little more variability, with statements and questions, possibly a few subordinate clauses. Phoenix is more natural than CSL-25k and Signum, although the language variability and complexity are modest. Indeed, it is safe to assume that the live interpretation can be influenced by the original speech, especially in terms of syntax, and will make little use of the structures typical of SL like iconicity and space (Section 3.1).

2.4. Annotation and performance metric

The three aforementioned corpora all share the same annotation scheme: for each SL sequence, the annotation Y_{CLSR} consists in the sequence of elicited or observed lexical signs[1]:

$$Y_{\text{CLSR}} = [g_1, \ldots, g_N], g_i \in \mathscr{G} \quad (1)$$

where $\mathscr{G} = \{g^1, \ldots, g^G\}$ is a dictionary of lexical signs, and N is the number of annotated lexical signs in the sequence. A straightforward performance metric for recognition is then the word error rate (WER), also referred to as Levenshtein Distance, applied to the expected sequences of lexical sign glosses. WER measures the minimal number of insertions I, substitutions S and deletions D to turn the recognized sequence to the expected sequence of length N:

$$\text{WER} = (I + S + D)/N. \quad (2)$$

Table 2 summarizes the WER achieved by best recent CLSR models on Signum, Phoenix and CSL-25k[2].

Recognizing the sequence of produced lexical signs in a SL utterance has undeniable values. For instance, it can help getting a grasp of the general topic of a SL discourse. However, as we will discuss in the next section, because of specificities of SLs including iconicity and the use of space, more natural corpora with finer annotations are needed to get closer to automatic SL understanding.

[1]It is to be noted that temporal information is lost in this annotation scheme. For Phoenix only, Koller et al. (2017) released estimated frame alignments from a hybrid model.

[2]On the CSL-25k corpus, Pu et al. (2019) have also considered a signer-independent dataset split, but all test sequences are seen in training. As this is formally equivalent to the problem of recognizing isolated gestures, we did not include their results in the table.

	Signum		Phoenix		CSL-25k	
	SD	SI	SD	SI	SD	SI
Koller et al. (2017)	4.8%	-	26.8%	44.1%	-	-
Cui et al. (2019)	2.8%	-	22.9%	39.8%	-	-
Pu et al. (2019)	-	-	36.7%	-	32.7%	-

Table 2: Word Error Rates of most recent lexical sign recognition models on three prevalent SL corpora, with signer-dependent (SD) and -independent (SI) settings.

3. Continuous *Sign Language* Recognition: a better consideration for linguistics

In this section, rather than a thorough description of the linguistics of SLs, we want to highlight some fundamental properties, arguing for a necessary redefinition of the CSLR problem with appropriate corpora.

3.1. Fundamental linguistic properties

3.1.1. Simultaneity

Although SL has often been described as a hand-articulated language, the linguistic role of non-manual articulators – including facial expressions, eye gaze, mouth, and body posture (Baker and Padden, 1978) – is actually as relevant as that of manual ones.

Notably, this great number of articulators make it possible to convey various information simultaneously (Vermeerbergen et al., 2007). This is illustrated on the SL sequence of Fig. 1, where expert annotations are given below video thumbnails (see Section 3.2.2 for more detail on annotation categories). Indeed, on frames 7, 8 and 9 of the sequence, the left hand represents part of a previously instantiated building, the right hand locates several restaurants while the facial expression insists on their important number.

3.1.2. Iconicity and visual grammar

Too often overlooked in the field of CSLR, iconicity is nonetheless a major SL feature. For Cuxac (2000), iconicity even has a structuring role in the linguistics of SL: building on the visual modality, it enables to *show while saying*. Using the signing space in a visual way to structure discourse is also fundamental, and forms the core of the visual grammar of SL.

Johnston and De Beuzeville (2014) draw a distinction between *Fully Lexical Signs* (FLS) and *Partially Lexical Signs* (PLS). In this classification, PLS include Pointing signs (PT), Fragment buoys and Depicting Signs (DS) (see Fig. 1 and associated Table 3 for a detailed example). DS are sometimes referred to as classifier constructions, classifier signs or illustrative signs. Sometimes building on purely lexical signs, they use proforms[3] to visually describe the location, motion, size, shape or the action of referrents, along with trajectories in the signing space.

As one can notice on the annotations of Fig. 1, a SL utterance can be mostly made of illustrative structures.

[3]Often referred to as classifiers. They are standard hand shapes used to represent a variety of common entities (Collomb et al., 2018).

Figure 1: LSF utterance from Dicta-Sign–LSF–v2 (Belissen et al., 2019), with a predominant use of space and iconicity (video reference: `S7_T2_A10` – duration: 4 seconds). From top to bottom: thumbnails, detailed annotation for the manual activity: fully lexical signs (FLS) and partially lexical signs (PLS), each on three tracks (right handed (RH), two handed (2H), left handed (LH)). More detail is given in Table 3 and Section 3.2.2.
Possible translation: *At the very center of this area, there is a large building surrounded by restaurants*.

Frame	Linguistic analysis of the manual activity
1, 2	**Depicting sign** construction, with the right hand localizing an area at the middle of the signing space, while the left hand helps representing its limit in space.
3	**Pointing sign** to the middle of the area in question. The left hand is static and maintains a fragment of the previous sign for spatial coherence, which is called a **fragment buoy**.
4	**Lexical sign** "Middle/center", insisting on the fact that what is going to be said is at the *very* center of the area.
5	**Depicting sign** representing the shape of a building, with a facial expression highlighting its massive size and central position.
6	The left hand has a **fragment buoy** function, from the building sign at the center of the setting. The right hand produces a one-handed version of the **lexical sign** "Restaurant" (its standard form is two-handed).
7, 8, 9	A standard **classifier** that can be understood as a smaller building is successively placed all around the area. Three instances are placed, but the face expression suggest that there are many of them, probably more than just three. The left hand still maintains the **reference point** to the large building.

Table 3: This table is a linguistic description of the manual activity in the SL sequence shown on Fig. 1, including *lexicon*, *buoys*, *proforms*, *pointing*, *iconic structures* and *spatial structure*.

3.2. Alternative corpora

Conversely to the corpora presented in Section 2.2, NCSLGR (Neidle and Vogler, 2012, ASL) and Dicta-Sign–LSF–v2 (Belissen et al., 2019, LSF) are two public corpora made of or including very natural SL and frame-aligned annotation on lexical and non-lexical levels.

3.2.1. NCSLGR

NCSLGR includes two categories of discourse. Most videos are made of elicited utterances, similar to that of Signum. However, the corpus also includes spontaneous short stories, with a lot more language variability.

Manual activity is annotated on two fields, one for the dominant hand and the other for the non-dominant hand. Annotations follow the conventions from Baker and Cokely (1980) and Smith et al. (1988), with: *lexical sign glosses*, *fingerspelling*, *hold signs* (hand position held at the end of a sign, not necessarily with a linguistic function), *pointing signs*, *depicting signs* (7 categories) with proforms and *gestures*. Non-manual activity is also annotated, with head movement and eye gaze among others.

3.2.2. Dicta-Sign–LSF–v2

Dicta-Sign–LSF–v2 is a public remake of the French Sign Language (LSF) part of the Dicta-Sign Corpus (Matthes et al., 2012), with cleaned and reliable annotations. The corpus is based on dialogue with very loose elicitation guidelines, it is thus highly representative of natural SL. The annotated manual activity is inspired from the convention of Johnston and De Beuzeville (2014), with:

- Fully Lexical Signs (FLS) on three tracks (dominant hand, two-handed, non-dominant hand):
- Partially Lexical Signs (PLS) on three tracks (dominant hand, two-handed, non-dominant hand):
  - Depicting Signs with proforms, under 7 types:
    location (of an entity, DS-L),
    motion (of an entity, DS-M),

size and shape (of an entity, DS-SS),
ground (spatial or temporal reference, DS-G),
action (handling of an entity, DS-A),
trajectory (in signing space, DS-T),
deformation (of a standard lexical sign, DS-X)

  - Pointing signs (PT)
  - Fragment buoys

- Non Lexical Signs (NLS), with fingerspelling, numbering and gestures.

Constructed actions, also referred to as role shifts or personal transfers were not annotated, even though they share some of the properties of DS.

As the two underlying linguistic models are different, the annotations do not follow the same conventions. However, one will notice that the difference is not so significant. With spontaneous SL and fine annotations on lexical and non-lexical levels, these two corpora pave the way for a newer and broader acceptation of CSLR.

In the next section, we discuss appropriate metrics and a possible formalization for CSLR that could include FLS, PLS and NLS.

3.3. C*SLR*: formalization

Let us consider a CSLR system dealing with M different linguistic descriptors $d^i, i \in \{1, \dots, M\}$, such that the annotation for a sequence of length T can be written as:

$$Y_{\text{CSLR}} = \begin{bmatrix} d^1 \\ \vdots \\ d^M \end{bmatrix} = \begin{bmatrix} d_1^1 & \cdots & d_T^1 \\ \vdots & \ddots & \vdots \\ d_1^M & \cdots & d_T^M \end{bmatrix}. \tag{3}$$

Each of these descriptors can be binary, categorical or continuous, depending on the encoded information. For instance, d^1 could encode recognized lexical signs (categorical), d^2 the presence/absence of a pointing sign (binary), etc. They could also include spatial information.

For a general CSLR model, each descriptor d^i must be assigned a specific performance metric. For a categorical descriptor like the temporal recognition of lexical signs, the accuracy Acc defined as the ratio of correctly labeled frames over the total number of frames T looks like a good candidate. For binary outputs like the presence/absence of a depicting sign, we have found frame-wise F1-score to be an informative metric. From the count of true/false positives/negatives, F1-score is defined as the geometric mean of precision P and recall R, that is:

$$\text{F1} = 2\left(P^{-1} + R^{-1}\right)^{-1}. \tag{4}$$

However, it is important to realize that even very good prediction models may not get close to F1 $= 1$. Amongst many reasons is the fact that the beginning and end of any linguistic phenomenon can be difficult to assess with precision. For very short realizations, a discrepancy of 1-2 frames at the beginning and end between predictions and annotations may worsen the score dramatically.

In order to reduce the impact that the subjectivity of the temporal localization of signs can have on the performance measure, true/false positives/negatives and thus F1-score can be evaluated on sliding windows as opposed to frame-wise. For instance in Belissen et al. (2020), all metrics are computed on a sliding window of four seconds length. Although F1-score is very informative, whether frame-wise or computed on sliding windows, a better performance metric is still to be engineered.

In Section 4, we analyze the recognition results of a first CSLR attempt on the depicting signs of Dicta-Sign–LSF–v2, and discuss the relevance and interest of this analysis.

4. Recognizing depicting signs in Dicta-Sign–LSF–v2

Belissen et al. (2020) developed a modern learning framework for the recognition of many linguistic descriptors. A simple representation of a signer is obtained by separately processing the head, body pose and hand shapes from any SL RGB video. A convolutional and recurrent neural network is then built on top of this representation, and trained to recognize lexical signs, depicting signs and others in a supervised learning fashion.

4.1. Analyzing a few sequences

In this section, we return to this work and focus on the recognition of depicting signs (DS) on Dicta-Sign–LSF–v2, with a finer discussion on the prediction of the trained model. Fig. 2a, Fig. 2b and Fig. 2c are three excerpts from one of the test videos of Dicta-Sign–LSF–v2. For each sequence, along with video thumbnails are given:

- Model predictions (dashed lines) compared to annotations (full lines) for the recognition of the broad category "Depicting signs" (F1-score in the caption),
- Fine annotations of the manual activity annotated on three tracks (right handed (RH), two handed (2H), left handed (LH)), both for fully lexical signs (FLS) and partially lexical signs (PLS).

The selected excerpts show good prediction performance, with F1-scores between 49% and 86%:

Fig. 2a The two depicting signs are almost perfectly recognized in this sequence, even though one will notice that F1-score is only 86%, due to slight temporal shifts.

Fig. 2b The unique depicting sign is detected, although the prediction lasts longer, lowering the F1-score to 62%. As a matter of fact, the previous sign ("Eiffel Tower") and the next one ("Visit") are somehow included in the illustrative setting so that it could make sense to recognize iconicity outside the annotated depicting sign of motion type. Close to frame 100, a form of constructed action could be recognized, even though not annotated.

Fig. 2c The only annotated depicting sign is recognized, although the F1-score is quite low at 49%: between frames 20 and 35, the models recognizes somethings that could look like unannotated constructed action.

4.2. Benefitting from this analysis

Based on these three different examples, a first analysis suggests that some *false positives* could actually make sense. Indeed, the relevance of a clear separation between lexical and illustrative levels has been discussed for a long time (Cuxac, 2000). A finely annotated corpus like Dicta-Sign–LSF–v2 could enable researchers to extend our work and question the relevance of prevalent linguistic descriptions of SLs. Conversely, the usual SLR setting, with lexical annotations and WER metric prevents one from conducting this type of research. It implicitly uses the hypothesis that SL discourse can be described with sequences of lexical signs, which we have shown is far from sufficient.

Highlighting the subjectivity in the annotation, these examples show that an appropriate metric is still to be designed.

Finally, annotation for constructed action would have been a great help for the analysis of the results, so it might be a future addition to this corpus. Indeed, the results might suggest that constructed action and depicting signs also have a lot in common.

5. Conclusion and perspectives

In this paper, we have insisted on the central role of iconicity and spatial structure in Sign Language discourse, highlighting the fact that Lexical Sign Recognition is only a part of the Continuous Sign Language Recognition task.

Since prevalent SL corpora have intrinsic limits in terms of generalizability and do not include annotations outside lexicon, we felt it was important to point out that richer corpora do exist, with fine temporal annotations.

As a first attempt on the French Sign Language corpus Dicta-Sign–LSF–v2, we have trained a recurrent neural network to recognize depicting signs. While noting the limits of F1-score as a metric, model performance was carefully analyzed. Decent scores are met, especially when considering the unclear boundary between lexical and depicting signs. Indeed, this frontier is dependent upon the chosen linguistic model, with no clear consensus on the matter.

Beside more analysis on the performance metric and linguistic model, future work will include further reflection on the ways spatial information can be annotated and included in automatic recognition models. On a long-term basis, we will also reflect on how to go from the detection of important discourse elements like illustrative structures to global Sign Language Understanding.

6. Bibliographical References

Baker, C. and Cokely, D. (1980). American Sign Language. *A Teacher's Resource Text on Grammar and Culture. Silver Spring, MD: TJ Publ.*

Baker, C. and Padden, C. (1978). Focusing on the nonmanual components of American Sign Language. Understanding language through sign language research.

Belissen, V., Gouiffès, M., and Braffort, A. (2020). Dicta-Sign-LSF-v2: Remake of a Continuous French Sign Language Dialogue Corpus and a First Baseline for Automatic Sign Language Processing. In *LREC*.

Collomb, A., Braffort, A., and Kahane, S. (2018). L'anatomie du proforme en langue des signes française:. quand il sert à introduire des entités dans le discours. *TIPA. Travaux interdisciplinaires sur la parole et le langage*, (34).

Cui, R., Liu, H., and Zhang, C. (2019). A Deep Neural Framework for Continuous Sign Language Recognition by Iterative Training. *IEEE Transactions on Multimedia*.

Cuxac, C. (2000). *La langue des signes française (LSF): les voies de l'iconicité*. Number 15-16. Ophrys.

Forster, J., Schmidt, C., Koller, O., Bellgardt, M., and Ney, H. (2014). Extensions of the Sign Language Recognition and Translation Corpus RWTH-PHOENIX-Weather. In *LREC*, pages 1911–1916.

Huang, J., Zhou, W., Zhang, Q., Li, H., and Li, W. (2018). Video-based Sign Language Recognition without Temporal Segmentation. In *32nd AAAI Conference on Artificial Intelligence*.

Johnston, T. and De Beuzeville, L. (2014). Auslan Corpus Annotation Guidelines. *Centre for Language Sciences, Department of Linguistics, Macquarie University*.

Johnston, T. and Schembri, A. (2007). *Australian Sign Language (Auslan): An Introduction to Sign Language Linguistics*. Cambridge University Press.

Koller, O., Zargaran, S., and Ney, H. (2017). Re-Sign: Re-Aligned End-to-End Sequence Modelling with Deep Recurrent CNN-HMMs. In *CVPR*, Honolulu, HI, USA, July.

Matthes, S., Hanke, T., Regen, A., Storz, J., Worseck, S., Efthimiou, E., Dimou, N., Braffort, A., Glauert, J., and Safar, E. (2012). Dicta-Sign – Building a Multilingual Sign Language Corpus. In *5th Workshop on the Representation and Processing of Sign Languages: Interactions Between Corpus and Lexicon (LREC 2012)*, pages 117–122.

Pu, J., Zhou, W., and Li, H. (2019). Iterative alignment network for continuous sign language recognition. In *CVPR*, pages 4165–4174.

Smith, C., Lentz, E., and Mikos, K. (1988). Vista American Sign Language series: Signing naturally.

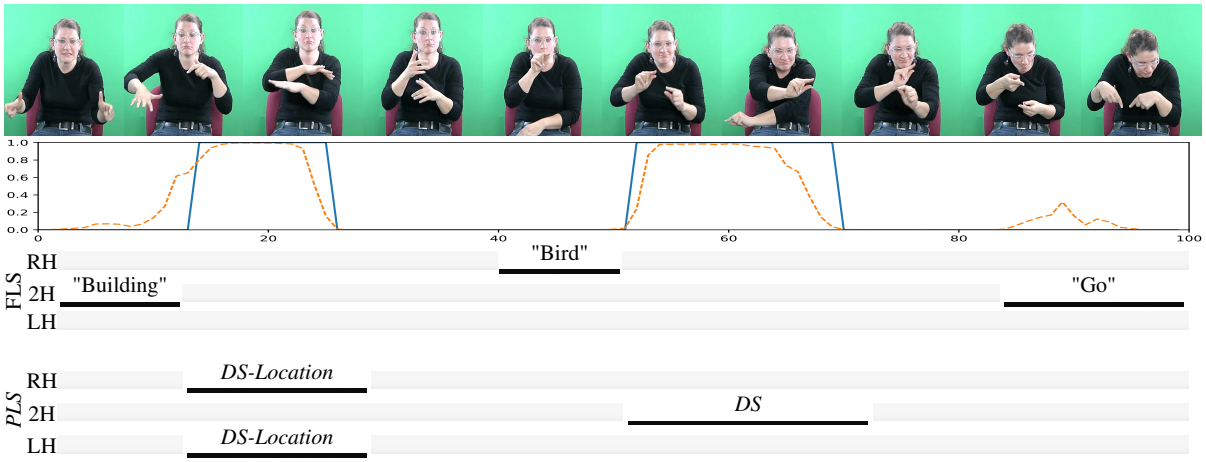
Vermeerbergen, M., Leeson, L., and Crasborn, O. (2007). *Simultaneity in Signed Languages: Form and Function*. Amsterdam studies in the theory and history of linguistic science. John Benjamins.

Von Agris, U. and Kraiss, K.-F. (2007). Towards a Video Corpus for Signer-Independent Continuous Sign Language Recognition. *Gesture in Human-Computer Interaction and Simulation*.

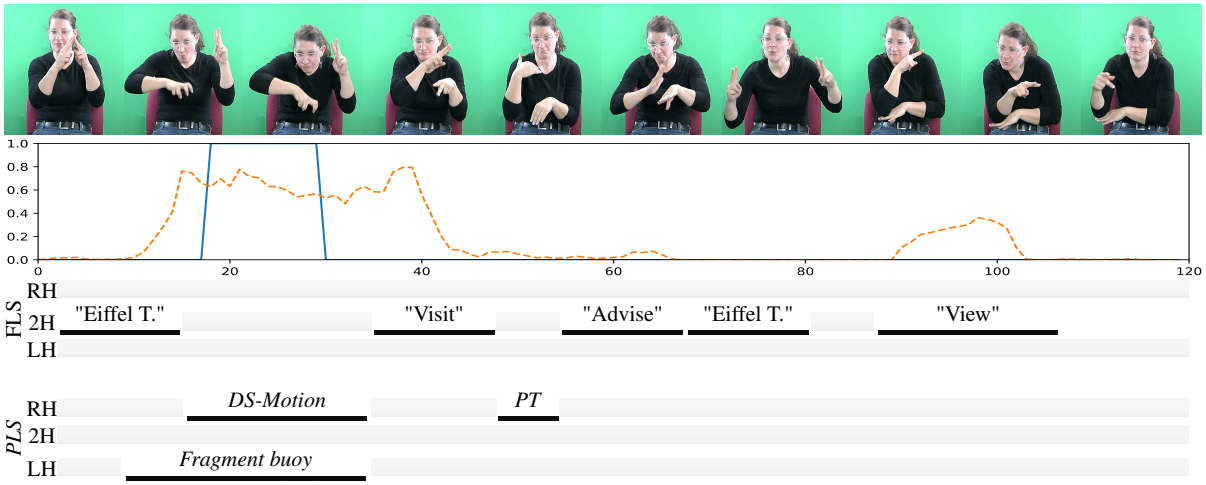
7. Language Resource References

Belissen, Valentin and Braffort, Annelies and Gouiffès, Michèle. (2019). *Dicta-Sign-LSF-v2*. Limsi, distributed via ORTOLANG (Open Resources and TOols for LANGuage), `https://www.ortolang.fr/market/item/dicta-sign-lsf-v2`, Limsi resources, 1.0, ISLRN 442-418-132-318-7.

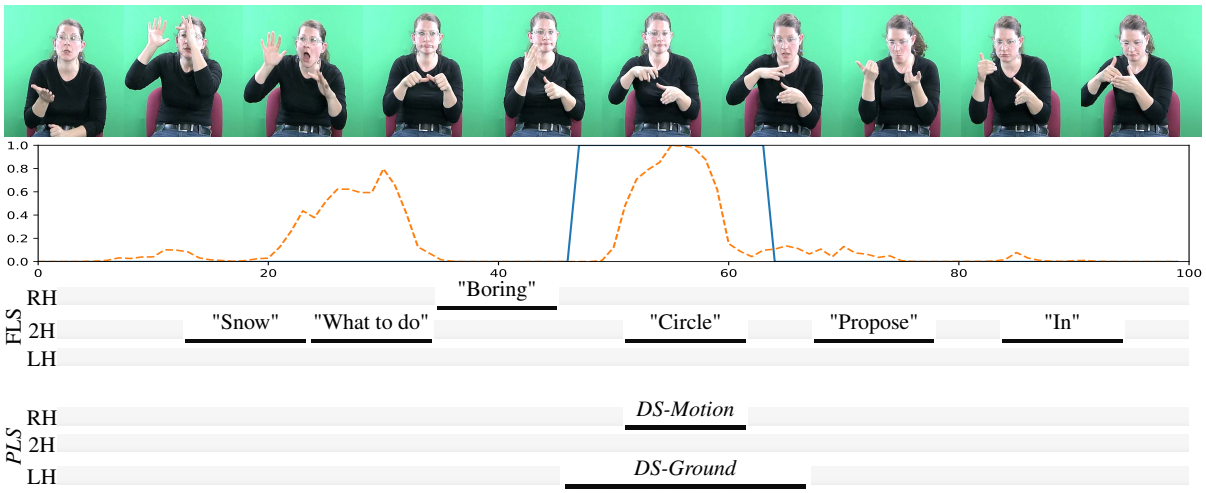
Neidle, Carol and Vogler, Christian. (2012). *NCSLGR*. American Sign Language Linguistic Research Project, `http://www.bu.edu/asllrp/ncslgr.html`.



(a) F1-score: 86% – *You should definitely go see this place where birds fly all around buildings.*



(b) F1-score: 62% – *I advise you to climb up the Eiffel Tower, you will then get a very nice panoramic view.*



(c) F1-score: 49% – *If you are stuck inside because of snow and you get bored, here is what I propose you to do.*

Figure 2: Three excerpts from Dicta-Sign–LSF–v2 (video reference: `S7_T2_A10`). For both sequences, from top to bottom: thumbnails, ground truth (solid) and predictions (dashed) for the recognition of Depicting signs, detailed annotation for the manual activity: fully lexical signs (FLS) and partially lexical signs (PLS), each on three tracks (right handed (RH), two handed (2H), left handed (LH)). Frame-wise F1-score is indicated in the caption, next to a proposed translation.

Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 13–20
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

Utterance-Unit Annotation for the JSL Dialogue Corpus: Toward a Multimodal Approach to Corpus Linguistics

Mayumi Bono [1&2], Rui Sakaida [1], Tomohiro Okada [2], Yusuke Miyao [3]
[1] National Institute of Informatics, [2] SOKENDAI (The Graduate University for Advanced Studies),
[3] The University of Tokyo
[1&2], 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, JAPAN
[3], 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, JAPAN
{bono, lui, tokada-deaf}@nii.ac.jp, yusuke@is.s.u-tokyo.ac.jp

Abstract

This paper describes a method for annotating the Japanese Sign Language (JSL) dialogue corpus. We developed a way to identify interactional boundaries and define a 'utterance unit' in sign language using various multimodal features accompanying signing. The utterance unit is an original concept for segmenting and annotating sign language dialogue referring to signer's native sense from the perspectives of Conversation Analysis (CA) and Interaction Studies. First of all, we postulated that we should identify a fundamental concept of interaction-specific unit for understanding interactional mechanisms, such as turn-taking (Sacks *et al.* 1974), in sign-language social interactions. Obviously, it does should not relying on a spoken language writing system for storing signings in corpora and making translations. We believe that there are two kinds of possible applications for utterance units: one is to develop corpus linguistics research for both signed and spoken corpora; the other is to build an informatics system that includes, but is not limited to, a machine translation system for sign languages.

Keywords: utterance unit, annotation, sign language dialogue

1. Introduction

This paper describes a method for annotating the Japanese Sign Language (JSL) dialogue corpus (Bono *et al*., 2014)[1]. Some linguists, including Deaf researchers who are interested in collecting sign language dialogue, began collecting data in April 2011. When we started, the general purpose of the project was to increase awareness of sign language as a distinct language in Japan. However, the academic aspects of the study recently became clear through interdisciplinary collaboration with engineering researchers, *i.e*., for natural language processing and image processing. In this paper, we introduce a preliminary result of our annotation process and annotated data, while explaining the concept of a 'utterance unit.' We anticipate that this concept will serve as a theoretical benchmark for promoting interdisciplinary research using spontaneous dialogue data in the corpus linguistics of sign languages.

2. Research Question and Background

In this study, we sought to find a way to identify interactional boundaries in sign languages and defined an utterance unit using various multimodal features accompanying signing.

2.1 Utterance Unit

The concept of utterance unit was already provided to segmenting and annotating spontaneous Japanese dialogues (Den *et al.* 2010; Maruyama *et al.,* in print). They propose a way of annotating utterance unit in two levels by emerging four linguistic and phonetic schemes, inter-pausal units, intonation-units, clause-units and pragmatic units.

In this paper, we define the concept of utterance unit for segmenting and annotating JSL dialogue data. We utilize JSL signer's native sense which is related to not only grammatical features but also multimodal features, such as mouth movements, non-manual movements, and gaze directions, to identify utterance unit. The method is based on classic observations in a research field of Conversation Analysis (CA) and Interaction Studies for spoken social interactions.

2.2 Sentence Unit

The previous studies on sign language linguistic have been focus on 'sentence unit' from the perspective of traditional linguistics. Crasborn (2007) introduces the workshop organized by his colleague and himself, which focuses on how to recognize a sentence in sign languages. He concludes that "we need to be alert to the risk of letting translations in another language influence our segmentation of signed language discourse, and keep our minds open for possible constructions that are modality specific" (Crasborn, 2007: 108).

Obviously, it should not rely on the writing system of spoken languages, because there is a risk of detecting an interactional chunk as a candidate of utterance unit using grammatical boundary of translated texts (e.g. JSL to Japanese). As widely known, there are some functional and grammatical utterance-final particles in Japanese, such as *ne* (ね), *yo* (よ), *yone* (よね) etc., they are possibly a signal of identifying interactional boundary. On the other hands, there is no functional and grammatical manual signs in JSL. In case of sign languages, these kinds of utterance final elements are spread in multimodal way, such as facial expressions and body postures.

2.3 Turn Constructional Units (TCUs) in CA

First of all, we had to introduce a classic concept of interaction-specific unit for understanding interactional mechanisms, such as turn-taking (Sacks et al., 1974). Conversation analysis (CA) is a sociological approach to the study of social interactions that applies the concept of turn constructional units (TCUs) (Sacks *et al.*, 1974) as fundamental building chunks of 'turns' in spoken

[1] Bono *et al*. (2014) introduces JSL colloquial corpus composed by dialogue part and lexicon part. Because we treat only dialogue part in this paper, we call it JSL dialogue corpus.

interactions, composed of utterances, clauses, phrases, and single words. CA research indicates that participants can anticipate TCUs and possible completion points of the ongoing turn using grammatical, prosodic, and pragmatic features of turn endings. Consequently, the turns in an interaction are exchanged smoothly among participants without difficulty.

Signers also naturally identify the boundaries of an utterance in social interaction, namely TCUs, to exchange turns visually. Signers probably recognize visual signals that are related to the grammatical, prosodic, and pragmatic completion points of turns. The concepts of TCUs and utterance units are similar. Here, we try to define an utterance unit in sign languages that aligns with the theoretical background of TCUs.

2.4 Applications

After identifying utterance units, we believe that they will have two applications: one is to develop corpus linguistics research for signed and spoken corpora; the other is to build an informatics system that includes, but is not limited to, a machine translation system for sign languages.

With regard to the former application, we anticipate that the research target of sign language studies will change drastically from example-based data to naturally occurring data, to study not only the grammatical aspects but also the social aspects of sign language interactions, such as turn-taking systems (Sacks *et al.*, 1974) and repair sequences (Schegloff *et al*., 1977) from the perspective of CA.

With regard to the latter application, we anticipate technical and theoretical breakthroughs in data collection and data storing using informatics technology, such as processing natural language and images. To recognize small hand and body movements in sign languages using image processing techniques (e.g., OpenPose), we will need to redesign the settings for data collection, lighting, frame rate, etc. If we want to translate sign language dialogue into spoken and written languages using deep learning or artificial intelligence technology, we will need to build a shared corpus to develop these systems.

The basic concept of the utterance unit is simple. However, we believe that it is a fundamental issue for developing sign language studies by combining research issues in linguistics and informatics.

3. Data

We collected JSL dialogues from 2012 to 2016. We have collected dialogues in 7 of the 47 Japanese prefectures (Table 1).

3.1 The first stage of data collection

As the first stage of data collection, we recorded videos of 40 deaf subjects in Gunma and Nara Prefectures (yellow in Fig. 1) from May to July 2012. Each prefecture has one school for the deaf. We obtained data from an age-balanced sample of individuals aged 30–70 years in each prefecture, and each age group was divided into same-sex pairs. Our participants from Gunma and Nara were in their 30s, 40s, 50s, 60s, and 70s, and included both male and female pairs.

3.2 The methods used for collecting dialogues

We used three methods to collect data: *interviews*, in which field workers and the assistants of native signers living in the same area who knew the procedures in advance asked the participants about their language, life, environment, etc. (for introductory purposes only, not open access); *animation narrative (AniN)*, in which one participant had memorized the story "Canary Row" and explained it to the other participant; and *lexical elicitation*, in which participants showed the corresponding signs for 100 slides of pictures and texts shown on a monitor, which is called JSL lexicon corpus (not included in this paper).

We collected pre-formed, lexical-level signing produced in a single-narrative setting and in spontaneous, utterance-level signing in a dialogue setting. In the single-narrative setting, we tried to detect enriched, deaf-specific signings using a theme for the narrative (*i.e*., folklore) and stimuli (pictures, images, etc.) to elicit signing at a lexical level. In a dialogue setting, we used video material to evoke a depictive signing (*i.e*., constructed action; Cormier, 2013) narrative task. We did not prepare a script for signing in advance. Consequently, the boundaries of the utterances were free, and were determined by participants who organized a turn-taking system in dialogue.

3.3 The amount of data

In the second stage of data collection, we collected data in Nagasaki, Fukuoka, Toyama, Ishikawa, and Ibaragi Prefectures, from 2014 to 2016 (green in Fig. 1). In this collection, we added two more dialogue tasks: '*my curry recipe (Cur)*' and '*proud of my country (Pro)*.'

Figure 1: Prefectures where dialogues were collected.

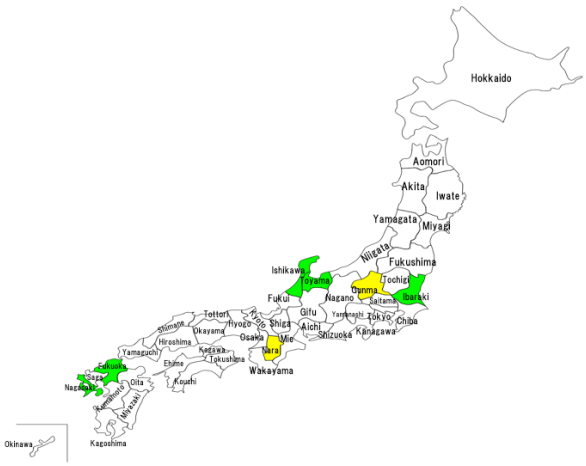


Table 1: Fundamental information of the dataset (collection year, number of dialogues, gender, and age range for each prefecture)

Prefecture	***Year***	***No. of Dialogue***	***Gender***	***Age Range***
Gunma	2012	10	M:10, W:10	30s-70s
Nara	2012	10	M:10, W:10	30s-70s
Nagasaki	2014	8	M:8, W:8	30s-70s
Fukuoka	2015	8	M:8, W:8	40s-80s
Toyama	2015	8	M:8, W:8	30s-70s
Ishikawa	2015	7	M:14, W:0	20s-80s
Ibaragi	2016	9	M:8, W:10	30s-70s
Total		60	120	20s-80s

Table 2: The percentage of video clips started putting basic annotations (word glosses and/or utterance unit glosses)

Prefecture	*AniN*	*Cur*	*Pro*	*Total*
Gunma	3/10			3/10
Nara	0/10			0/10
Nagasaki	8/8	8/8	4/8	20/24
Fukuoka	8/8	8/8	8/8	24/24
Toyama	8/8	8/8	4/8	20/24
Ishikawa	7/7	7/7	4/7	18/21
Ibaragi	0/9	0/9	0/9	0/27
Total	34/60 (56%)	31/40 (77%)	20/40 (50%)	85/140 (60%)

The total number of participants is 120 in 60 dialogues (See Table 1). The total recording time in corpus is 40 hours 52 minutes and 28 seconds. In the case in which we narrow down only dialogue tasks, AniN, Cur, and Pro, the total recording time is 15 hours 42 minutes and 59 seconds. As you can see in Table 2, the total number of video clips is 140. We have started putting basic annotations, word glosses and/or utterance unit glosses, to 85 files (60%). Actually, the annotated number of tokens, nearly equal to word gloss, is 27,371, November 26, 2019. Three independent video clips, collected camera A, B and C, were synchronized using Final Cut Pro. The original combined-angles image includes the interlocutor's back recorded by cameras B and C; there also is dead space, shown in black in Fig. 2. Cropped combined-angles images do not include the interlocutor's back and there is no dead space. The video images from all camera angles were enlarged to facilitate detailed analysis.2

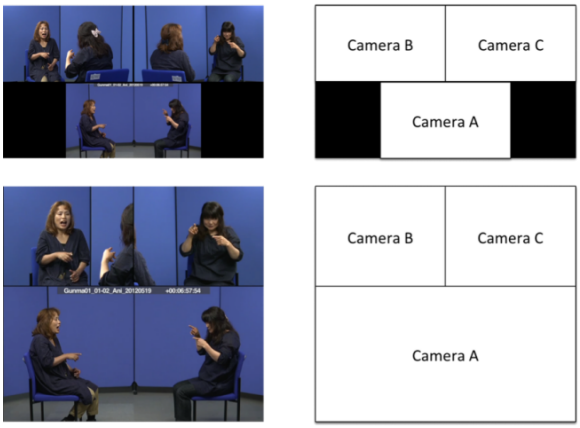


Figure 2: Image of two versions of the three camera angles: original (top), cropped (bottom).

4. Utterance Unit Annotation

We set two annotation levels: individual and integrated levels. The individual level is composed of four tiers: word gloss, mouth movement, non-manual movement (NMM), and gaze tiers. If annotators find features that can be used to a define utterance unit, such as the narrator's nodding behavior at turn-endings or a gaze shift from the signing space to the interlocutor, they classified them into each tier. Note that the annotated information does not include everything that happened in a dialogue; annotators tagged only information related to the grammatical, prosodic, or pragmatic features of turn endings.

At the integration level, all information annotated at individual levels is combined to identify utterance units. Figure 3 presents an example of the tier structure. 'NS_11_SH_40F' is the participant's information, which in this case means a female in her 40s who comes from the southern part (SH) of Nagasaki (NS), participant ID 11. Each tier has the participant's information to avoid confusing the annotated data among annotators. The tier names are placed after the participant information.

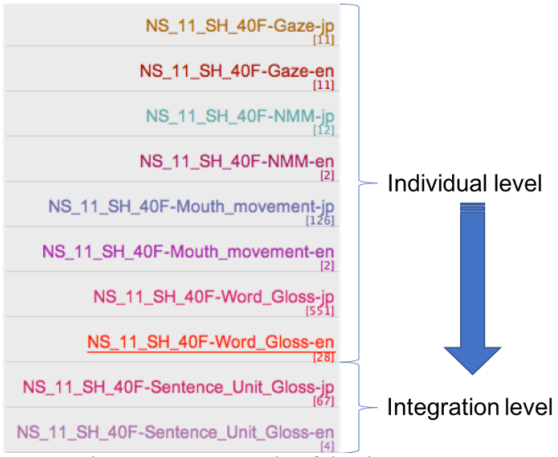


Figure 3: An example of the tier structure.

4.1 Individual Level

As mentioned above, the individual level is composed of four tiers. However, only the word gloss tier is mandatory; the others are optional.

4.1.1 Word Gloss Tier (mandatory)

Because JSL still does not have a digital dictionary or ID gloss system, such as *SignBank* and *Global Sign Bank*, the annotators placed the Japanese and English meanings of the signed word in the word gloss tier directly.

Before annotating the word gloss, all annotators learned the concept of the gesture unit (GU) proposed by Kendon (1970, 2004) to identify the start and end points of signed words.

One of our original plans was to establish a physical and hand movement unit smaller than the word gloss (Bono *et al.*, 2014). We applied the concept of the GU to annotate the beginning and end points of signed words. The GU is the interval between successive rests of the limbs, rest positions, or home positions. A GU consists of one or several gesture phrases. A gesture phrase is what we

[2] For more information about the JSL Dialogue Corpus, http://research.nii.ac.jp/jsl-corpus/research/data/manual/manual.html

intuitively call a 'gesture,' which consists of up to five phases: preparation, stroke, retraction, and pre- and post-stroke hold phases. We used the preparation, stroke, and retraction phases to identify the start and end points of a word gloss.

Figure 4 presents an image of a word unit for word gloss annotation. When several sign words form one utterance, such as in the lower image, the retraction phase is replaced by the preparation phase of the next signed word. In utterance unit annotation, annotators did not include the

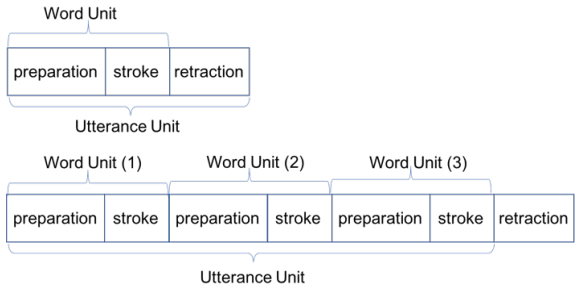


Figure 4: An image of word units for word gloss annotation.

GU phases, although they learned about this concept before annotating the word gloss.

Annotators write the meanings of signed words in capital letters on ELAN. If one signed word has a meaning that is a combination of two or more spoken words, such as TRY HARD (Excerpt 1, Section 5.1), it is connected by '.', *e.g.*, 'TRY.HARD', to indicate that it is one word in sign. If it is repeated several times, the number of repeats follows, *e.g.*, 'TRY.HARD (rep3)'.

4.1.2 Mouth Movement Tier (optional)

As in sign linguistics, there are two kinds of mouth movement in sign languages: mouth gestures and mouthing. A mouth gesture has a grammatical function, such as being adverbial to hand signings. Mouthing refers to shapes and movements that originate from spoken languages and are used while signing. Annotators classify these mouth movements by noting 'mg:' for mouth gesture and 'm:' for mouthing at the beginning of every annotation. When signers use mouth movements at the same time as signing, annotators place the following hand signing information in parentheses, *e.g.*, 'TRY.HARD(rep3) (m: ga-m-ba-ri-ma-su [try hard])' (Excerpt 1).

As mentioned above, we do not annotate all mouth movements, but only those related to utterance units, *i.e.*, the beginning or end of turns.

4.1.3 NMM Tier (optional)

We included a tier for non-manual movement (NMM) for grammatical and ungrammatical elements made by the body. For instance, a signer who has the role of a narrator may use nodding to show utterance endings. We would classify 'nod' into this tier.

4.1.4 Gaze Tier (optional)

Kendon (1967) observed the systematic mechanism of gaze direction at the ends of turns, including TCU endings and transition relevance places (TRPs) from the perspective of CA, from a psychological perspective. A shifting gaze can be crucial for identifying an utterance boundary, such as when signers shift their attention from the signing space to the interlocutor at the beginning or end of an utterance to confirm the interlocutor's understanding of the narrative. We would classify gaze directions into this tier.

4.2 Integration Level

Glosses at the individual level are combined at the integration level, which is the utterance unit. Annotators make a general judgement of the start and end of an utterance using information from individual levels annotated in advance.

All tiers in both levels are annotated by two native Deaf signers and one CODA (Children of Deaf Adults). Currently, these three annotators annotated five dialogues

Data ID	Gender of pair	Age	Task	Prefecture	Length of dialogue	(1) No. of Word Unit Gloss	(2) No. of Utterance Unit Gloss	Words in Utterance (1)/(2)
Data 1	Male	60's	*Cur*	Toyama	0:09:44	499	102	4.89
						624	119	5.24
Data 2	Female	60's	*Cur*	Toyama	0:08:04	304	38	8.00
						483	64	7.55
Data 3	Female	40's	*Cur*	Toyama	0:07:09	358	67	5.34
						490	70	7.00
Data 4	Female	40's	*AniN*	Toyama	0:10:04	896	109	8.22
						258	70	3.69
Data 5	Female	40's	*AniN*	Nagasaki	0:05:55	551	67	8.22
						57	33	1.73
Total	M:1; F:4	40's; 60's	*Cur*:3; *AniN*:2	Toyama:4, Nagasaki: 1	0:40:56	4,520	739	

Table 2: Results of Annotations.

on ELAN as a first test. As you can see in Table 2, total length of targeted five dialogues is almost 41 min. The average of tokens per min. is about 113[3].
Furthermore, table 2 shows the number of words gloss, the number of utterance unit gloss, and words per utterance.
Data 1, 2, and 3 are dialogues conducted the task of my curry recipe, and data 4 and 5 are dialogues conducted the task of animation narrative (Canary Row). There is a difference of the frequencies of utterance unit gloss between these tasks. In curry recipe, the number of words in utterance between participants in dialogues are balance such as 4.89 and 5.24 in data 1, on the other hand, in animation narrative, those are unbalanced, such as 8.22 and 1.73 in data 5. In Animation narrative task, there is the tendency that the participant who watched movie clip in advance holds turns and have multiple TCU in a turn, and the intelocuter gives small number of words to narrator in short reactions, such as, *uhn hm, I see* in English.

5. Analysis

In the following analyses, we present three excerpts analyzed using the CA framework to clarify how we integrate the features in tiers at the individual level to identify utterance units.

5.1 Excerpt 1: TRPs with Mouthing

In excerpt 1, signer TY_12 (lower tiers in ELAN of excerpt1, Figure 5) says, "*You should cook a delicious meal for your husband.* (HUSBAND/ FOR/ DELICIOUS/ MAKE/ GIVE (m: a-ge-te [give])". Signer TY_11 answers by mouthing and signing, "*Yes, I'll do my best.* ((m: ha-i [yes])/ TRY.HARD (rep3) (m: ga-m-ba-ri-ma-su [try hard]))".

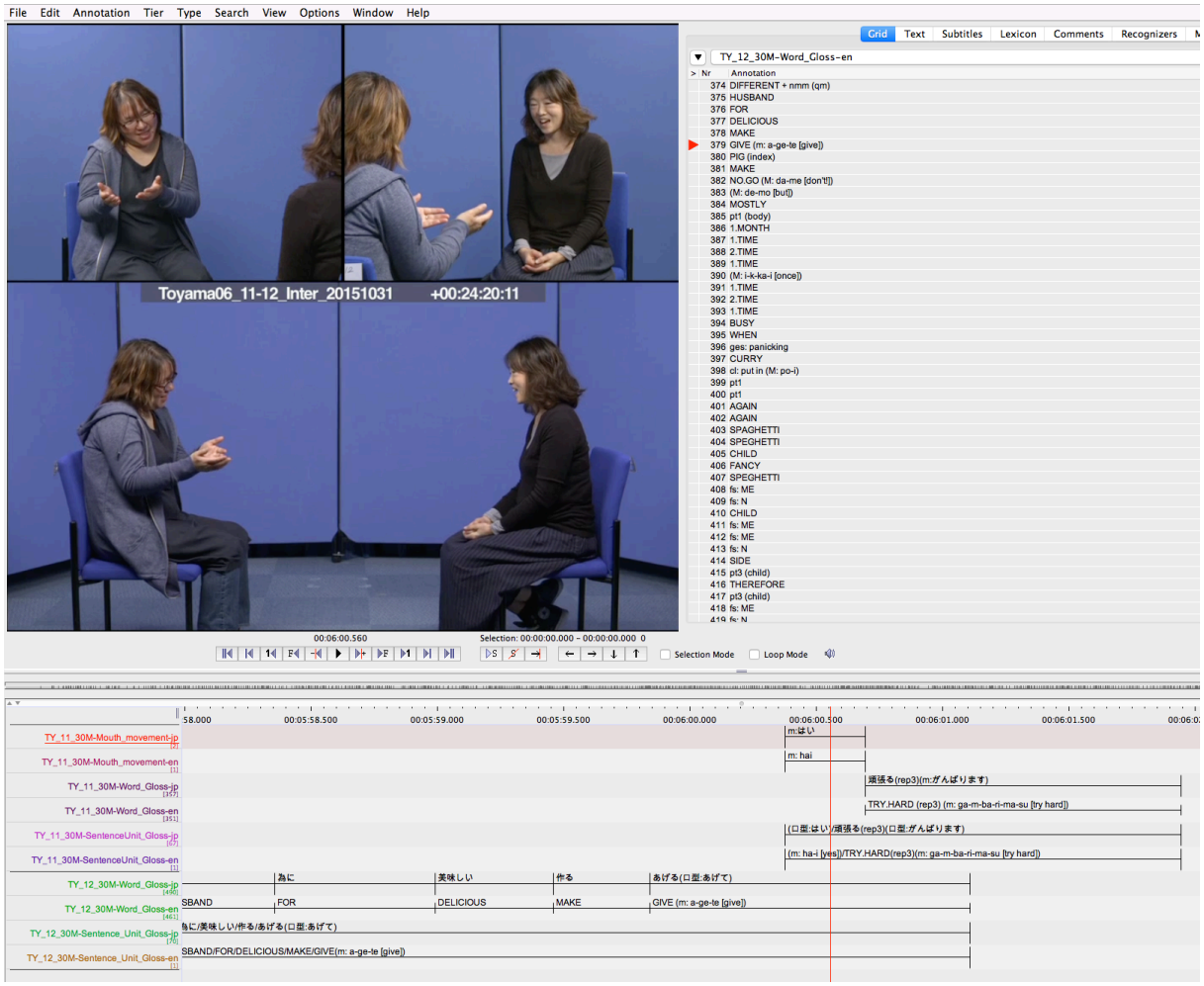


Figure 5: ELAN annotation of Excerpt 1, Transition relevance place with mouthing.

[3] Reviewer #1 pointed out to us that the average of tokens per min in a dialogue setting is about 120 in DGS (German Sign Language). There is a similar tendency in our corpus.

Interestingly, TY_11 produces mouthing before hand signing to answer TY_12's recommendation. Specifically, during the middle of the final sign /GIVE/, TY_12 starts her answer by mouthing 'ha-i [yes]'. /GIVE/ is a subsidiary verb in Japanese and is also an agreement verb in JSL. TY_12 moves both of her hands to the right, where the semantic meaning "TY_11's husband" was given in advance (Figure 5) (Liddell, 2012). Before signing /GIVE/, TY_12's utterance is almost grammatically and semantically completed. Consequently, within /GIVE/, specifically close to the end of the stroke for /GIVE/, is the earliest sequential position for TY_11 to give a response. And TY_11 gives a first response, not by signing, but by mouthing 'm: ha-i [yes]'.

In this segment, their utterances overlap. We assume that this is a typical case of transition relevance place (TRP) in sign language dialogues. It implies that we should include mouthing when discussing the utterance units.

5.2 Excerpt 2: Narrative and Role-shift with Gaze

Next, we discuss utterance completion and the narrator's gaze behavior. In excerpt 2, signer NS_11 (upper tiers in the ELAN of excerpt 2, Figure 6) is telling a story about the animation she has watched, called 'Canary Row' which is an animation clip used in Gesture Studies (e.g. McNeill, 1992).

She describes a famous first scene of it by producing multiple utterances: (1) "*A chick is swinging inside a bird cage*. (cl:human:un(stop)/ cl:human:chick:having a swing(stop)/ cl:sphere(circle)/ cl:un (something that swings like a swing in the sphere)/ cl:the shape of a cage/)"; (2) "*The cute chick is playing on the swing.* (pt3(rep)/ CUTE/ cl:human:chick: (ges:flaps the wings)/ cl:human:chick: (ges: enjoying playing on a swing)"; and (3) "*A cat finds the chick, and climbs something like a pillar quickly*. (pt1 (meaning: pt (cat))/ CAT/ cl:human: cat: (ges: looks around, notices something and claps his hands)/ cl:human:cat:climbs something like a pillar(stop)/ cl:explanation of something like a pillar standing upright/cl:human:cat (ges:looks around quickly) /cl: human: cat climbs the pillar quickly /cl:climbs/ cl:climbs+NMM)".

As we can see, she uses lexical expressions only for /CUTE/ and /CAT/ in this part. Moreover, these lexical signs are accompanied by mouthing. Other expressions are depicting signs (cl) and gestures (ges) without mouthing. This style of signing is very familiar in sign language narrative talk.

We focus on the narrator's gaze behavior at the boundary of each utterance. In each ending of all the utterances, the narrator (NS_11) looks at the interlocutor (NS_12). Furthermore, the narrator gives a nod at the end of utterances (1) and (3) and the interlocutor gives a response, such as /UNDERSTAND/ at these points.

The analysis of excerpt 2 revealed that the gaze directions and head nods accompanying the narrative provide clues for the interlocutors to identify utterance units. Moreover, the integration of these clues shows the utterance boundary more strongly.

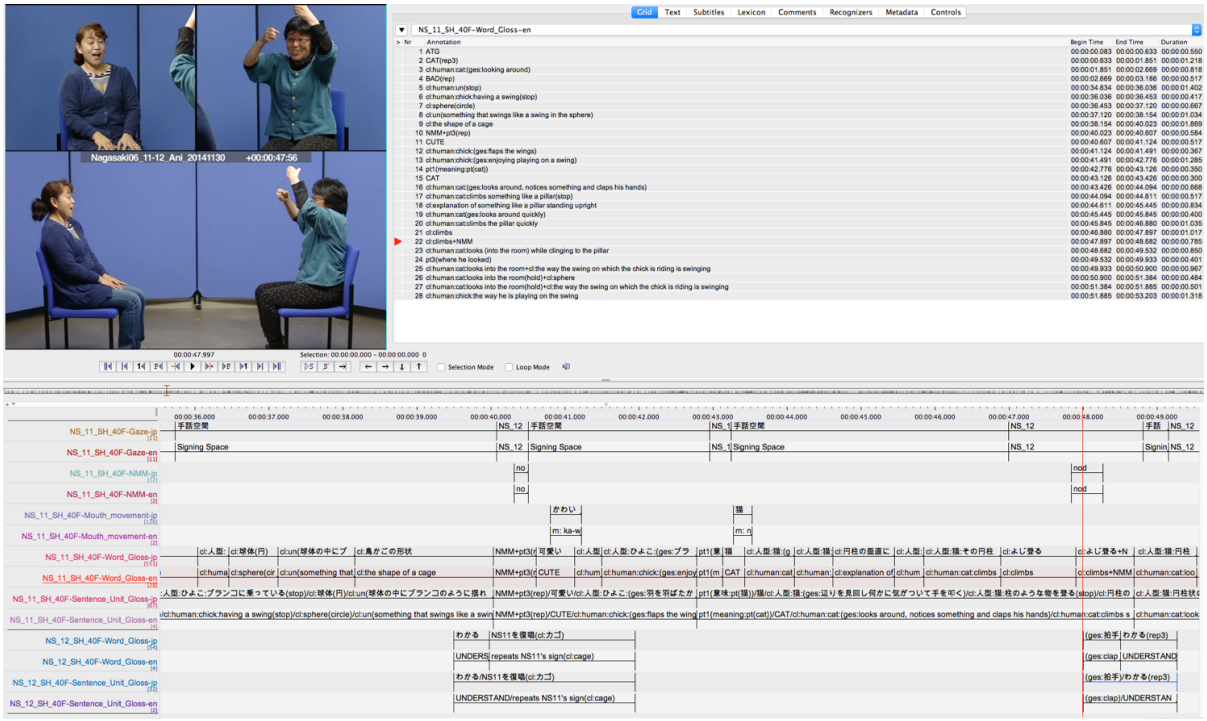


Figure 6: ELAN annotation of Excerpt 2, Narrative and role-shift with gaze.

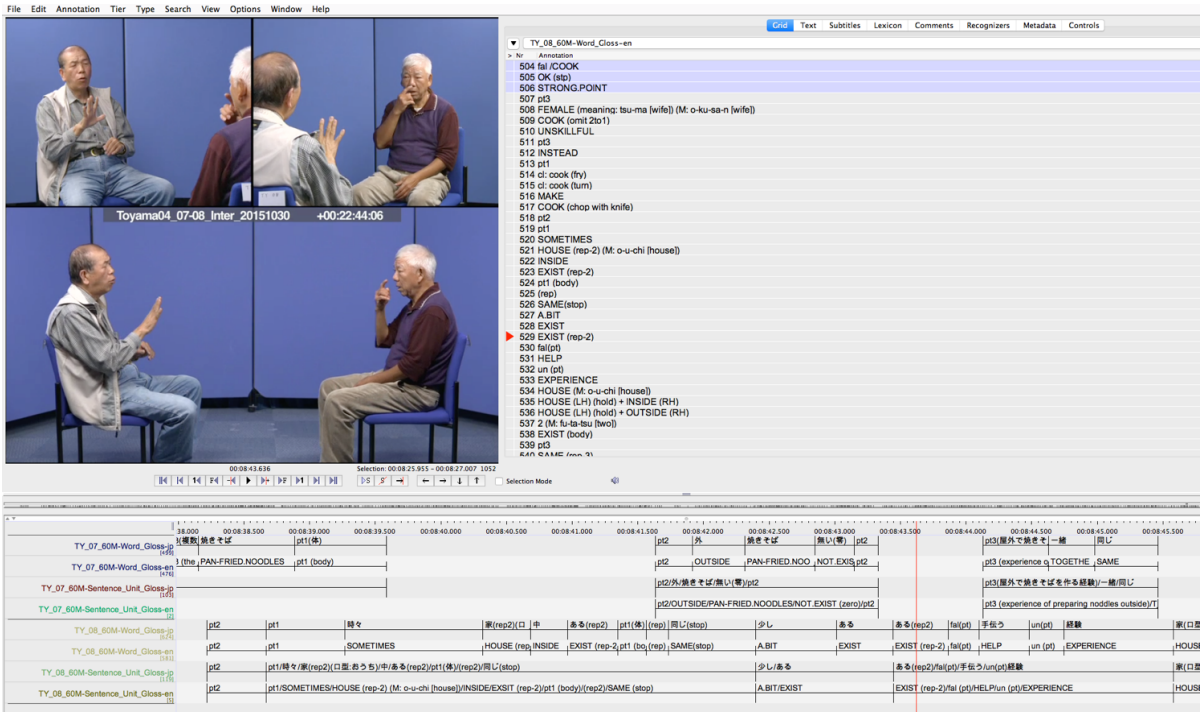


Figure 7: ELAN annotation of Excerpt 3, Other-initiated repair with overlap.

5.3 Excerpt 3: Other-initiated Repair with Overlap

Finally, excerpt 3 is an example in which they exchange their utterances orienting to turn-taking system, not a narrative or one-way signing. Excerpts 2 and 3 appear similar at the point where the interlocutor's actions – responses in excerpt 2 and confirmation questions in excerpt 3 – overlapped with the current signer's signing and are retroactively defined as utterance units.

In excerpt 3, TY_08 starts to explain his experience cooking pan-fried noodles (*yakisoba*) by producing sequential multiple utterances, (1) "*I'll make it in my house*. (pt1/ SOMETIMES/ HOUSE (rep-2) (M: o-u-chi [house])/ INSIDE/ EXIST (rep-2)/ pt1 (body)/ (rep2)/ SAME (stop))", (2) "*Sometimes…* (A.BIT/EXIST)", and (3) "*Yeah, I helped cooking*. (EXIST (rep-2)/ fal (pt)/ HELP/ un (pt)/ EXPERIENCE)".

However, these three utterances are not connected like the narrative in excerpt 2. From TY_07's questions, we can see how TY_08's multiple sequential utterances are connected. That is, TY_07 asks TY_08 a question to display his understanding (Sacks, 1992). TY_08 answers him as soon as possible, as in the utterance (2) mentioned above.

To obtain more detail, although TU_08 continues his turn with the utterance (1), TY_07 asks him, "*Don't you make them outside* (like camping)? (pt2/ OUTSIDE/ PAN-FRIED.NOODLES/ NOT.EXIST (zero)/ pt2)" during the utterance's final particles, pt1(rep), which is a sandwich construction with pt1 in the utterance-onset, which is a TRP. Then, TY_08 gives an answer by connecting his utterance with his previous utterance, as in utterance (2), "*Sometimes…*". That is the earliest place for him to give a response.

The exchanges in excerpt 3 are related to the concept of other-initiated repair sequence (Schegloff, 1977). There are four techniques for others to initiate repair: open class forms, category-specific interrogatives, repeats of the trouble-source turn, and candidate understandings (Sidnell & Stivers, 2013). A TY_07's question overlapping an utterance (1) is used as pt2 (pointing at the interlocutor) at the onset and offset of an utterance, which makes it clear that there is a something trouble for TY_07 in TY_08's utterance, who was pointed out by pt2.

Consequently, the annotators segmented TY_08's signing into three parts: utterances (1), (2), and (3). Utterance (1) is a description of his experience; utterance (2) is an answer to TY_07's question; and utterance (3) is an elaboration of his own answer in utterance (2).

From the analysis of excerpt 2, we found that not only annotated multimodal features in tiers at the individual level, but also the sequential structure of dialogues, are clues used to identify utterance units.

6. Conclusions

This paper describes an annotation method for the Japanese Sign Language (JSL) dialogue corpus (Bono *et al.*, 2014) by defining the concept of an utterance unit. By analyzing three excerpts, we showed how complicated it is to identify utterance units using a combination of signing and various other features. However, annotators who are all native signers (Deaf and Coda) with a native understanding of JSL

used multimodal features to identify the utterance units. We found that it was very difficult to establish a standard criterion for finding features among annotators.

The utterance is a fundamental unit in languages. It is obvious that we cannot rely on the written system of spoken languages. Defining the utterance unit in sign languages will have useful applications, such as setting a fundamental unit for storing data in sign language corpora, and for manual or machine translation using advanced technologies.

7. Bibliographical References

Bono, M., Kikuchi, K., Cibulka, M. and Osugi, Y. (2014). Colloquial corpus of Japanese Sign Language: A design of language resources for observing sign language conversations. Proc. of the ninth International Conference on Language Resources and Evaluation Conference, pp.1898-1904.

Crasborn, O. (2007). How to recognise a sentence when you see one. *Sign Language & Linguistics* **10-2**, pp.103-111.

Den, Y., Koiso, H., Maruyama, T. and Yoshida, N. (2010). Two-level Annotation of Utterance-units in Japanese Dialogs: An Empirically Emerged Scheme. Proc. of Seventh International Conference on Language Resources and Evaluation, pp.17-23.

Kendon, A. (1967). Some functions of gaze direction in social interaction. *Acta Psychologica*, **26**, pp.22-63.

Kendon, A. (1970). Movement coordination in social interaction. *Acta Psychologica*, **32**, pp.1-25.

Kendon, A. (2004). *Gesture: Visible Action as Utterance*. Cambridge: Cambridge University Press.

Liddell, S. (2012). *Gesture, Grammar and Meaning in American Sign Language*. Cambridge University Press.

Maruyama, T., Den, Y. and Koiso, H. (in print). Design and annotation of two-level utterance units in Japanese. In Izre'el, S. *et al*. (eds) *Search of Basic Units of Spoken Language: A corpus-driven approach.* John Benjamins.

McNeill, D. (1992). *Hand and mind*. Chicago, IL: University of Chicago Press.

Sacks, H., Schegloff, E. A. and Jefferson, G. (1974). A Simplest Systematics for the Organisation of Turn-Taking for Conversation, *Language*, **50**, pp.696–735.

Sacks, H. (1992). *Lectures on Conversation, Volumes I and II,* Edited by G. Jefferson with Introduction by E.A. Schegloff, Blackwell, Oxford.

Schegloff, E. A., Jefferson, G. and Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation, *Language*, **53**(2), pp. 361-382.

Sidnell, J. and Stivers, T. (2013). *The handbook of conversation analysis*. Wiley-Blackwell.

8. Appendix

HUSBAND	A sign is indicated in upper case.
NOT.EXIST (zero)	When there are more than two expressions for what is being signed, the expression selected is indicated in parentheses ( ).
cl:the shape of a cage	Classifiers or depicting signs are indicated in lower case.
(m: a-ge-te [give])	Mouthing. Hyphens (-) are used to delimit each kana syllable. The translation of the mouthing is added within brackets [ ].
pt1	Pointing to speaker him/herself
pt2	Pointing to hearer
pt3	Pointing to neither speaker nor hearer
pt1(body)	The object pointed to is indicated in lower case within parentheses ( ).
(ges:flaps the wings)	Gestures
NMM	Non manual markers
(rep3)	Reduplications. When the number of iterations is known, it is indicated as "(rep2)" for two iterations, "(rep3)" for three and so forth.
(stop)	A cut-off or truncation
un	Unclear hand movements
fal	Signing errors
(meaning: pt (cat))	The meaning of a sign in the conversational context is sometimes described.

Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 21–26
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

Measuring Lexical Similarity across Sign Languages in Global Signbank

Carl Börstell[1], **Onno Crasborn**[1], **Lori Whynot**[2]
[1]Radboud University / [2]Northeastern University
Erasmusplein 1, 6525 HT Nijmegen, The Netherlands / 360 Huntington Ave., Boston, MA 02115, USA
c.borstell@let.ru.nl, o.crasborn@let.ru.nl, l.whynot@northeastern.edu

Abstract

Lexicostatistics is the main method used in previous work measuring linguistic distances between sign languages. As a method, it disregards any possible structural/grammatical similarity, instead focusing exclusively on lexical items, but it is time consuming as it requires some comparable phonological coding (i.e. *form* description) as well as concept matching (i.e. *meaning* description) of signs across the sign languages to be compared. In this paper, we present a novel approach for measuring lexical similarity across any two sign languages using the Global Signbank platform, a lexical database of uniformly coded signs. The method involves a feature-by-feature comparison of all matched phonological features. This method can be used in two distinct ways: 1) automatically comparing the amount of lexical overlap between two sign languages (with a more detailed feature-description than previous lexicostatistical methods); 2) finding exact form-matches across languages that are either matched or mismatched in meaning (i.e. true or false friends). We show the feasability of this method by comparing three languages (datasets) in Global Signbank, and are currently expanding both the size of these three as well as the total number of datasets.

Keywords: lexical similarity, linguistic distance, cross-linguistic comparison, lexicostatistics, false friends, mutual intelligibility

1. Introduction

Glottolog 4.1 (Hammarström et al., 2019), one of the (if not *the*) most comprehensive language databases to date, lists 194 sign languages of the world. However, we know very little about the possible genealogical relationships between different sign languages, and many such claims are based solely on historical records of language contact and influences. Due to the scarcity of historical documentation and the fact that all sign languages should still be considered under-studied, little is known about linguistic distances between sign languages, a metric which could be used to estimate possible phylogenies. However, there are few methods for calculating linguistic distances that could be applied to sign languages, considering the format and quantity of available data. Previous work in this domain has mainly used *lexicostatistics*, a method of comparing form overlap between lexical items across languages based on concept lists with translations into the languages in question. For sign languages, such studies have mostly been undertaken on an areal basis, with the intention of using lexical overlap as a metric for the likelihood of two languages being related (Woodward, 1991; Woodward, 1993; Woodward, 2000; McKee and Kennedy, 2000; Guerra Currie et al., 2002; Johnston, 2003; Bickford, 2005; Al-Fityani and Padden, 2010). These studies have in common that they compare the form similarity of two signs with the same meaning (i.e. concept-matched) from different sign languages, although the exact method for comparing sign forms across languages has varied between studies. In general, these studies consider the four basic form parameters of a sign (see Figure 1) and count two forms with all parameter values equal as *identical*, forms with one parameter value differing as *similar*, and more differing values as *different* forms.[1]

Parameter	Value
Location	neutral space
Handshape	B hand
Orientation	palm forward
Movement	ipsilateral movement

Figure 1: The NGT sign NEE-E ('no') with form parameter descriptions (Crasborn et al., 2020b).

This methodology has proven valid in the sense of finding greater similarity across sign languages known to be related (Johnston, 2003), but it can also be a somewhat crude measure that finds similarity that is purely incidental, and some studies have thus either tried to include iconic motivation as an additional factor in such measures (Ebling et al., 2015), or introduced a more fine-grained method for comparing sign forms across languages by separating form parameters into more detailed (sub)features (Yu et al., 2018). Here, we follow a path more similar to the latter, by using the uniformly coded cross-linguistic sign language lexical database *Global Signbank* (Crasborn et al., 2020a) to automatically measure lexical similarity across sign languages.

An ultimate goal with this method is to predict communicative success in cross-signing contexts (Zeshan, 2015; Byun et al., 2018) and mutual intelligibility across sign languages (Sáfár et al., 2015). The hypothesis is that languages with similar phonologies may show overlap in sign forms, which may or may not encode the same meaning. If the meaning

[1]Some studies would conflate parameters and thus look at three rather than four parameters.

Language	Sign entries	Coded signs	% coded
NGT	4,026	3,531	88%
CSL	2,248	568	17%
IS	200	200	100%

Table 1: Language datasets and number of coded signs in Global Signbank (Crasborn et al., 2020a).

overlaps (true friends), the prediction is that mutual intelligibility is higher; if not (false friends), this could be an impeding factor for cross-signing. As an example, the NGT (*Nederlandse Gebarentaal*; Sign Language of the Netherlands) sign WAT-A ('what'; Figure 2a) is identical to the ASL (American Sign Language) sign WHERE, and the NGT sign WAAR-A ('where'; Figure 2b) is identical to the ASL sign WHAT. This overlap in form but mismatch in meaning may disrupt cross-signing, since the addressee recognizes the form but associates it with a different meaning. Disruption in comprehension due to these types of false friends were indeed found in a study on comprehension of International Sign (IS) for signers of Japanese Sign Language and Auslan (Australian Sign Language) (Whynot, 2015).

2. Data and Methodology

2.1. Global Signbank

In our current data, we have a number of sign languages stored in an online lexical database called *Global Signbank* (Crasborn et al., 2020a). The languages – each represented as a separate *dataset* – are accessed in a graphical user interface (Figure 3) in which signs can be searched by translation keywords (e.g. in Chinese, Dutch, English), sign glosses (unique labels for signs), and are displayed as video files (`.mp4`), animated images (`.gif`), and still images (`.png`), together with fields containing phonological form-descriptions of signs.

We use data from three languages, in order of size of the datasets (see Table 1): NGT – 4,026 signs; 3,531 (88%) of which have phonological coding (Crasborn et al., 2020b); Chinese Sign Language – CSL, Shanghai variety; 2,248 signs; 568 (17%) of which have phonological coding (Crasborn et al., 2020c); and International Sign – IS; 200 signs; 200 (100%) of which have phonological coding (Whynot, 2020). NGT and CSL are two urban, unrelated languages; IS is a sign system based on mainly European-derived sign languages, used primarily as a form of communication at international deaf events, not used as an L1 in any community (Hiddinga and Crasborn, 2011; Whynot, 2015).

The relevant form-description fields in Global Signbank included in our sign similarity comparison are listed below:

- Handedness
- Strong Hand
- Weak Hand
- Handshape Change
- Relation between Articulators
- Location
- Relative Orientation: Movement
- Relative Orientation: Location
- Orientation Change
- Contact Type
- Movement Shape
- Movement Direction
- Repeated Movement
- Alternating Movement

2.2. Concepticon

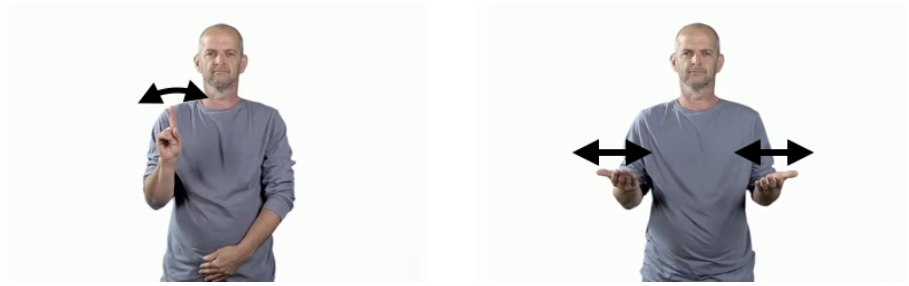
Since we want to look at lexical similarity across languages, we need a way to map form to meaning uniformly across datasets. We use the Concepticon concept list database (List et al., 2019) for this purpose. Concepticon is a database of collected concept lists from a diverse set of linguistic studies, compiled into one master list with links to individual lists collected – one list being the ECHO Swadesh list for sign languages (Woll et al., 2010). We use a crude method of mapping the English keywords/glosses in Global Signbank sign entries (see Figure 3) to concepts in Concepticon through string-matching. By doing so, we can compare signs not only from form to meaning (by manually looking at form-matches and evaluating their meaning-correspondence), but also meaning to form (by comparing those forms that are mapped to the same concept). The matching, mapping, and comparison steps are described in the following section.

2.3. Similarity Measure

After matching all language datasets to Concepticon as described above, we proceed to the automatic comparison of sign forms. Here, we compare any two signs on each form-description field and compute the number of overlapping fields. Since not all fields are relevant to all signs, we calculate the differences only for fields that have a value listed for both signs. This means that the comparison of a one-handed and a two-handed sign will result in a different value for the *Handedness* field (1 vs. 2 hands), but the field *Weak Hand* will be skipped altogether as it is only relevant for two-handed signs. Thus, we get a binary value (0 = different; 1 = same) for each relevant field, and divide the total by the number of fields compared to arrive at a sign similarity score between 0 and 1. The comparison is done with an automated script.

In the first step, we want to compare *all* signs in one language to *all* signs in another language. This means that we disregard meaning in this first automatic comparison stage, and let our script iterate through all possible sign pairs across datasets and store the similarity score for each such pair. This step of our cross-linguistic sign comparison is illustrated in Figure 4 using NGT and CSL.

In the second step, we want to compare only those pairs of signs across languages that are matched to the same Concepticon concepts. Since some concepts may be matched to several sign entries within a language dataset (due to form variations), the script iterates through each variant for a concept in one language and compares it to each variant for the same concept in the other language, and subsequently return the sign pair with the highest lexical similarity. This is illustrated schematically in Figure 5 in which only signs matched to the concept 'no' are compared to each other. In



(a) The NGT sign WAT-A ('what'). (b) The NGT sign WAAR-A ('where').

Figure 2: The NGT signs WAT-A (a) and WAAR-A (b) (Crasborn et al., 2020b).

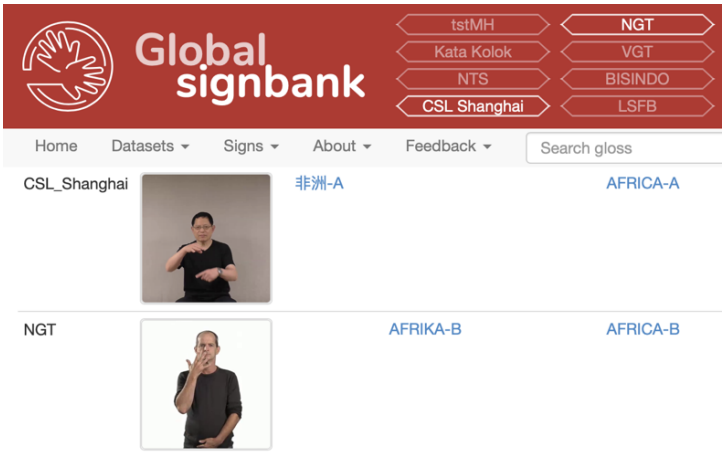


Figure 3: The graphical user interface of Global Signbank, showing the search results for *Africa* in two datasets (languages): CSL and NGT. Glosses are available in English for both datasets, as well as Chinese for CSL and Dutch for NGT.

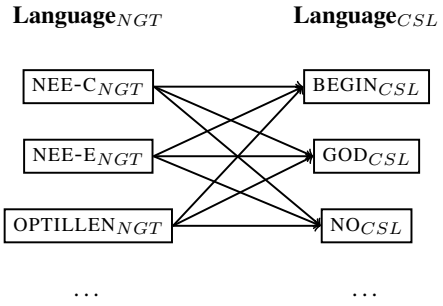


Figure 4: Cross-linguistic sign form comparison, all combinations.

this minimal example, the sign variants NEE-C ('no'; Figure 6a) and NEE-E ('no'; Figure 6b) in NGT are both compared to the CSL sign NO ('no'; Figure 6c), after which only the pair NEE-E$_{NGT}$ and NO$_{CSL}$ is kept as it has the highest degree of overlap (.88), only differing in the CSL sign having a repeated movement.

3. Results

In the first step, we compared all sign forms in one language dataset against all sign forms in another language

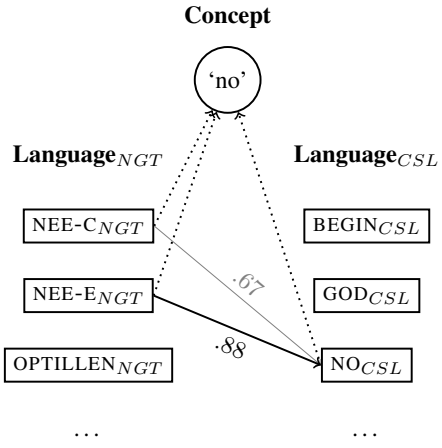


Figure 5: Cross-linguistic sign form comparison, only concept-matched combinations (dotted lines) compared. Highest similarity sign pair (thick black line) returned.

dataset, for each pairing across our three languages: NGT–CSL; NGT–IS; CSL–IS. Since the number of signs coded differs greatly across our three language datasets, the number of sign form matches are expected to differ accordingly. Indeed, we find most form overlaps with the pair-

(a) The NGT sign NEE-C ('no').

(b) The NGT sign NEE-E ('no').

(c) The CSL sign NO ('no').

Figure 6: The NGT signs NEE-C (a) and NEE-E (b) (Crasborn et al., 2020b), and CLS sign NO (c) (Crasborn et al., 2020c).

Pair	Matches	True	False
NGT–CSL	30 (5.3%)	12	18
NGT–IS	10 (5%)	6	4
CSL–IS	0 (–)	–	–

Table 2: Form-matches and number of true vs. false friends across all language pairings.

Pair	Matches	Mean	Median
NGT–CSL	194	.406	.375
NGT–IS	62	.436	.444
CSL–IS	43	.373	.333

Table 3: Concept-matches and their mean and median form-similarity scores across all language pairings.

ings that involve NGT – the largest of our datasets – and also more overlaps for the NGT–CSL pair than the NGT–IS pair, given that the CSL dataset is larger than the IS dataset. As shown in Table 2, 30 sign pairs are matched as form-identical across NGT and CSL. We look at each matched pair individually in order to evaluate whether they also match in meaning (true friends) or not (false friends). Of these 30 sign pairs, 12 pairs constitute true friends in that they have exact or similar meaning-matches: an example of an exact match in form and meaning is NGT and CSL signs for 'good' (Figures 8a–8b); an example of an exact form-match with a similar meaning is NGT JEZUS-A ('Jesus'; Figure 9a) and CSL GOD ('God'; Figure 9b). 18 pairs constitute false friends, sign pairs for which the forms are identical but the meanings are different: one example of this is the NGT sign OPTILLEN-A ('to lift'; Figure 10a) and the CSL sign BEGIN ('to begin'; Figure 10b).

For the NGT–IS pair, we find 10 sign pairs with identical forms, 6 of which are true friends and 4 of which are false friends, and for the CSL–IS pair we find no form-matches whatsoever. We find proportionally more true friends between NGT and IS than between NGT and CSL, which could be indicative of a general closer lexical similarity between the former languages than the latter. However, seeing as the datasets and the absolute numbers of form-matches are miniscule, this conclusion would be premature.

In the second step, we compared only those sign forms that were concept-matched to Concepticon. Again, we find more matches for the larger datasets, unsurprisingly as the number of potential matches is only as big as the smaller dataset (language) in any given pair (cf. Table 1). Thus, NGT–CSL has 194 concept-matched signs, NGT–IS has 62, and CSL–IS has 43. Concept-matched signs with mean and median similarity scores are shown in Table 3, and the distribution of similarity scores are shown in Figure 7.

These results point to NGT and IS generally having a higher

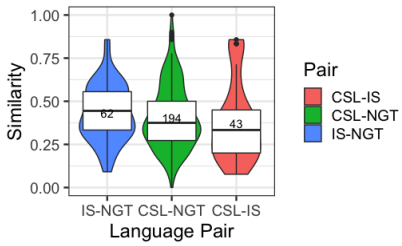


Figure 7: Distribution of sign form-similarity scores in concept-matched sign pairs across all language pairings.

similarity for signs denoting the same concept than either pairing including CSL. This, together with the higher proportion of true friends from the first step of the lexical comparison, may suggest a closer lexical distance between NGT and IS than any of the CSL pairings – and similar European vs. Asian sign language splits have been suggested (Yu et al., 2018). However, since our datasets are still small and also disproportionate in size, this is at best a preliminary suggestion in need of further examination.

4. Discussion

In this paper, we have described a method for comparing lexical similarity as an indicator of linguistic distance across sign languages represented as datasets in Global Signbank. Our method works in two directions: 1) from form to meaning (whether signs that overlap in form also overlap in meaning, i.e. are true or false friends); 2) from meaning to form (to what extent the phonological forms of signs for the same concept across languages are (dis)similar. With larger datasets (in terms of both languages and sign entries), we see the potential of this method to be used for lexicostatistics across a range of languages

(a) The NGT sign GOED-A (‘good’).

(b) The CSL sign GOOD (‘good’).

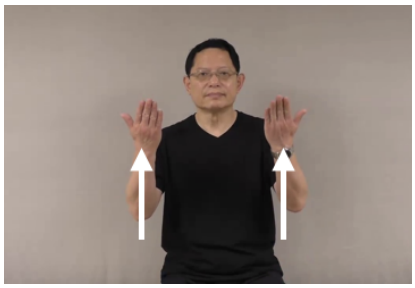
Figure 8: The NGT sign GOED-A (a) (Crasborn et al., 2020b) and CLS sign GOOD (b) (Crasborn et al., 2020c).

(a) The NGT sign JEZUS-A (‘Jesus’).

(b) The CSL sign GOD-A (‘God’).

Figure 9: The NGT sign JEZUS-A (a) (Crasborn et al., 2020b) and CLS sign GOD (b) (Crasborn et al., 2020c).

(a) The NGT sign OPTILLEN-A (‘to lift’).



(b) The CSL sign BEGIN (‘to begin’).

Figure 10: The NGT sign OPTILLEN-A (a) (Crasborn et al., 2020b) and CLS sign BEGIN (b) (Crasborn et al., 2020c).

through a (semi-)automated process, which would speed up the process compared to a purely manual comparison and allow for pairwise comparisons across a large set of languages which could be clustered along multiple dimensions (Bickford, 2005; Yu et al., 2018).

Furthermore, we hope to use some of these methods in order to quantify linguistic distances (focusing on lexical similarity) and apply the results to our ongoing project investigating cross-signing – that is, communication across different sign languages. When signers engage in cross-signing, they bring their individual sets of linguistic resources and skills, which include the use of material from their own primary language(s) as well as the adjustment and adaptation to the communicative context. Previous research has shown that deaf signers are able to communicate successfully without sharing any signed or spoken language, after only a short amount of time in the cross-signing context (Zeshan, 2015; Byun et al., 2018). Nonetheless, little is known about whether linguistic distance (as in a high degree of lexical similarity) influences the degree of communicative success in cross-signing contexts, though one could assume that cross-signing success is affected by lexical similarity, much like mutual intelligibility based on the amount of overlap in conventional lexical items (Sáfár et al., 2015). Such effects on comprehension have been shown in a study on IS, in which signers whose languages use signs similar to corresponding signs in IS would perform better on an IS lexical comprehension task (Whynot, 2015).

One of the unique features of the method outlined above is that it takes variation into account. Signers have in their linguistic repertoire not only their own preferred (e.g. dialectal, sociolectal) sign form for a concept, but are also familiar with other signs used in their language community. In our method, the best match sign pair is always used in cases of variants, which accounts for having passive knowledge of a sign form–meaning mapping without necessarily

producing it. In cross-signing interactions, these multiple variants constitute part of the communicative resources that a signer brings to the table, and in our measures of lexical similarity, we include this aspect of linguistic knowledge.
Using this method, we hope to establish a metric for linguistic distances not only for linguistic classification (in terms of lexical typology or genealogy), but also for the expected communicative success in cross-signing contexts. Historical connections between sign languages (based on, often scarce, historical records) may offer some explanation for potential cross-linguistic comprehension and mutual intelligibility. However, such cross-linguistic intelligibility may be possible without relatedness, by virtue of iconic motivation. That is, if the languages involved happen to recruit similar iconic patterns in sign formation, cross-signing comprehension may be more successful. Thus, although lexical form similarity is one metric that could easily be used to estimate cross-linguistic comprehension, including a more schematic perspective on iconicity mappings (Ebling et al., 2015) may prove to be necessary too.

5. Acknowledgements

This study was funded by the *Netherlands Organisation for Scientific Research* (NWO) grant number 277-70-014.
We would like to acknowledge our team: Tashi Bradford, Aurélia Nana Gassa Gonga, Maya de Wit, and Merel van Zuilen. A special thanks to Anique Schüller and Neil Ray for annotating Signbank datasets for this study.

6. Bibliographical References

Al-Fityani, K. and Padden, C. (2010). Sign languages in the Arab world. In Diane Brentari, editor, *Sign languages: A Cambridge language survey*, pages 433–450. Cambridge University Press, New York, NY.

Bickford, J. A. (2005). *The signed languages of Eastern Europe*. SIL International & University of North Dakota.

Byun, K.-S., de Vos, C., Bradford, A., Zeshan, U., and Levinson, S. C. (2018). First Encounters: Repair Sequences in Cross-Signing. *Topics in Cognitive Science*, 10(2):314–334.

Ebling, S., Konrad, R., Boyes Braem, P., and Langer, G. (2015). Factors to Consider When Making Lexical Comparisons of Sign Languages: Notes from an Ongoing Comparison of German Sign Language and Swiss German Sign Language. *Sign Language Studies*, 16(1):30–56.

Guerra Currie, A.-M. P., Meier, R. P., and Walters, K. (2002). A crosslinguistic examination of the lexicons of four signed languages. In Richard P. Meier, et al., editors, *Modality and structure in signed and spoken language*, pages 224–237. Cambridge University Press, Cambridge.

Hiddinga, A. and Crasborn, O. (2011). Signed languages and globalization. *Language in Society*, 40(4):483–505.

Johnston, T. (2003). BSL, Auslan and NZSL: Three signed languages or one? In Anne Baker, et al., editors, *Cross-linguistic perspectives in sign language research: Selected papers from TISLR 2000*, pages 47–69. Signum, Hamburg.

McKee, D. and Kennedy, G. (2000). Lexical comparisons of signs from American, Australian, British and New Zealand Sign Languages. In Karen Emmorey et al., editors, *The signs of language revisited: An anthology to honor Ursula Bellugi and Edward Klima*, pages 49–76. Lawrence Erlbaum Associates, Mahwah, NJ.

Sáfár, A., Meurant, L., Haesenne, T., Nauta, E., De Weerdt, D., and Ormel, E. (2015). Mutual intelligibility among the sign languages of Belgium and the Netherlands. *Linguistics*, 53(2):353–374.

Whynot, L. (2015). *Assessing comprehension of international sign lectures: Linguistic and sociolinguistic factors*. dissertation, Macquarie University.

Woodward, J. (1991). Sign language varieties in Costa Rica. *Sign Language Studies*, 73:329–346.

Woodward, J. (1993). The relationship of sign language varieties in India, Pakistan, and Nepal. *Sign Language Studies*, 78:15–22.

Woodward, J. (2000). Sign language and sign language families in Thailand and Viet Nam. In Karen Emmorey et al., editors, *The signs of language revisited: An anthology to honor Ursula Bellugi and Edward Klima*, pages 23–47. Lawrence Erlbaum Associates, Mahwah, NJ.

Yu, S., Geraci, C., and Abner, N. (2018). Sign Languages and the Online World Online Dictionaries & Lexicostatistics. In Nicoletta Calzolari, et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Zeshan, U. (2015). "Making meaning": Communication between sign language users without a shared language. *Cognitive Linguistics*, 26(2):211–260.

7. Language Resource References

Onno Crasborn, et al., editors. (2020a). *Global Signbank*. Radboud University, Nijmegen. https://signbank.science.ru.nl.

Crasborn, O., van der Kooij, E., Zwitserlood, I., and Ormel, E. (2020b). Nederlandse Gebarentaal (NGT) dataset in Global Signbank. In Onno Crasborn, et al., editors, *Global Signbank*. Radboud University, Nijmegen. https://signbank.science.ru.nl.

Crasborn, O., van Zuilen, M., and Gong, Q. (2020c). Chinese Sign Language (CSL) dataset in Global Signbank. In Onno Crasborn, et al., editors, *Global Signbank*. Radboud University/Fudan University, Nijmegen/Shanghai. https://signbank.science.ru.nl.

Hammarström, H., Forkel, R., and Haspelmath, M. (2019). Glottolog database 4.1. https://glottolog.org.

List, J.-M., Greenhill, S. J., Rzymski, C., Schweikhard, N. E., and Forkel, R. (2019). Concepticon 2.1. https://concepticon.clld.org.

Whynot, L. (2020). International Sign (IS) dataset [Whynot 2015] in Global Signbank. In Onno Crasborn, et al., editors, *Global Signbank*. Radboud University, Nijmegen. https://signbank.science.ru.nl.

Woll, B., Crasborn, O., van der Kooij, E., Mesch, J., and Bergman, B. (2010). *Extended Swadesh list for signed languages*. http://www.let.ru.nl/sign-lang/echo/.

Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 27–32
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

Optimised Preprocessing for Automatic Mouth Gesture Classification

Maren Brumm[1, 2], Rolf-Rainer Grigat[2]
[1]Institute of German Sign Language and Communication of the Deaf,
University of Hamburg, Gorch-Fock-Wall 7, 20354 Hamburg, Germany,
[2]Vision Systems, Hamburg University of Technology, Harburger Schloßstraße 20, 21079 Hamburg, Germany
maren.brumm@uni-hamburg.de, grigat@tuhh.de

Abstract

Mouth gestures are facial expressions in sign language, that do not refer to lip patterns of a spoken language. Research on this topic has been limited so far. The aim of this work is to automatically classify mouth gestures from video material by training a neural network. This could render time-consuming manual annotation unnecessary and help advance the field of automatic sign language translation. However, it is a challenging task due to the little data available as training material and the similarity of different mouth gesture classes. In this paper we focus on the preprocessing of the data, such as finding the area of the face important for mouth gesture recognition. Furthermore we analyse the duration of mouth gestures and determine the optimal length of video clips for classification. Our experiments show, that this can improve the classification results significantly and helps to reach a near human accuracy.

Keywords: Sign Language Recognition/Generation, Machine Translation, SpeechToSpeech Translation, Statistical and Machine Learning Methods

1. Introduction

Mouth gestures are facial expressions in the context of sign language, that do not refer to words of a spoken language. They are an important part of the German Sign Language that can be crucial for understanding the meaning of signing (Von Agris et al., 2008).
A corpus with annotated mouth gestures would be helpful for further research, but is very time-consuming to acquire. The aim of this work is to develop a method to automatically classify mouth gestures by training a neural network. This could eliminate time-consuming manual annotations as well as advance automatic sign language translation. However, mouth gesture classification is a challenging task even for humans. Some mouth gesture classes are very similar to each other and the style, duration and intensity of a mouth gesture can vary significantly from person to person. Another issue is the small size of training material available. Therefore, careful preprocessing of the data can significantly improve the results, as it helps to reduce the input data to the necessary information only. We evaluate the effect of the usage of different regions of interest (ROIs) within a frame, different methods to convert the videos to a fixed length, as well as different clip durations.
Earlier works on non-manuals use facial landmarks as features. (Neidle et al., 2014) detect non-manual grammatical markers and (Luzardo et al., 2014) estimate a mouth state (open / close / narrow /...) by geometric features based on facial landmarks.
More recent works often use neural networks. To our knowledge there are only two publications on automatic mouth gesture recognition (Wilson et al., 2019), (Brumm et al., 2019). We extend the work of (Brumm et al., 2019), however, without the use of an avatar. Our work is also similar to (Wilson et al., 2019) but we use a different neural network architecture.
There are some papers on the related subject of mouthing, which are facial expressions in sign language that do refer to spoken words, such as (Koller et al., 2014), (Koller et al., 2015) and more on the broader field of spoken word recognition and lip reading like (Chung and Zisserman, 2016), (Chung et al., 2017), (Afouras et al., 2018) and (Martinez et al., 2020). The architecture of the neural network used in this work is based on (Petridis et al., 2018), who use spatiotemporal convolution followed by a 34-layer ResNet and 2-layer BGRU.

2. Dataset

Our dataset was generated from the DGS corpus of the DGS-Korpus project at the University of Hamburg [1]. It consists of 4177 mouth gestures from 281 different signers appearing in natural conversation. We identified 21 classes of mouth gestures, that appear frequently in the corpus. They were annotated independently by two different annotators. The annotators also provided the exact start and end point of the mouth gestures within the video.
However, for some of the 21 mouth gesture classes we could not find a sufficient number of training examples and in some cases two of the mouth gesture classes are too similar to be differentiated with a reasonable accuracy even for human annotators. We therefore reduced the number of mouth gesture classes for automatic classification to ten, by combining very similar mouth gestures to one mouth gesture class and leaving out classes with less than 52 examples.
This results in a dataset with 2842 examples of ten different mouth gesture classes. The number of examples per class varies between 52 and 615. Table 1 describes the ten chosen classes and shows how many examples are in the dataset for each of them.
To estimate the accuracy with which humans can perform mouth gesture classification, we use the inter-annotator agreement of the two annotators. As the annotators were originally asked to classify the data to 21 different mouth

[1] https://dgs-korpus.de

gesture classes, we can not determine the exact human classification accuracy (or inter-annotator agreement) for our ten class classification problem. Considering only examples where both annotators give a class within the ten chosen classes, the accuracy is 79.13%. Considering all clips where the first annotator gives a class within the ten chosen classes the accuracy is 66.40%. The real human accuracy is somewhere in this range.

3. Neural Network Architecture

The architecture of the neural network used is based on the work of (Petridis et al., 2018). However, we only use the visual stream of their two-stream network. It consists of a spatiotemporal convolutional layer, a 34-layer ResNet and a 2-layer BGRU. The network was pretrained on the Lip Reading in the Wild (LRW) database (Chung and Zisserman, 2016).

4. Proposed Preprocessing Options

4.1. Region of Interest

The original videos show the upper body of the person as well as the background, as can be seen in figure 1a. The first step is therefore to extract the region of interest (ROI). Our aim is to make the ROI as small as possible without loosing relevant features. This is especially important as our dataset is small, which makes it more difficult for the network to distinguish between relevant and non-relevant artefacts in the image.
We consider three possible ROIs shown in figures 1c, 1d and 1e. The first is a close-up of the mouth. The second shows the lower part of the face, as helpful features may also be located on the cheeks or the nose. The third option is to use the whole face, to also include possible features located on the eyes, eyebrows and forehead.
Figure 1b shows the ROI that was used in (Petridis et al., 2018). As we use their pretrained model, similarity effects have to be taken into account, as described in section 5.3.

4.2. Frame Sampling

Naturally the mouth gestures differ in length. We consider two different methods to transform the mouth gesture clips to a fixed length.
The first is to up- or downsample the clips to the required number of frames, as described in (Wilson et al., 2019). If a clip is too long, frames are removed at even intervals. If it is too small, frames are doubled at even intervals. This assures, that the mouth gesture is visible from the start until the end. But frames in between might be missing or doubled.
The second option we propose, is to cut out a consecutive number of frames left and right of the midpoint of the mouth gesture. This may lead to parts of the mouth gesture being cut off, if the actual mouth gesture is longer than the number of frames used or other video material being included that is not part of the mouth gesture, if the mouth gesture is shorter. But the clip that is cut out is consecutive. Both methods require knowledge of the location of the mouth gesture in the video. However, this information is available during training only. When applying the method to unlabelled data the location of the mouth gesture can only be approximated by the location of the hand gesture accompanying the mouth gesture. In this case the continuous method might be advantageous as the network has learned to classify clips that are incomplete or show unrelated material.
Section 6.2 shows the comparison of the different sampling methods. The results on approximating the mouth gesture location are shown in section 6.3.

4.3. Clip Duration

The duration of the clips used as input to the neural network is an important parameter. If a clip is too short, parts of the mouth gesture are cut off. If it is too long, it shows facial actions not related to the mouth gesture. Both aspects could lead to poorer training results. This is especially true if the exact timing of the mouth gesture is unknown and the starting and end point is determined by the hand gesture that is accompanying the mouth gesture, which would be the case in a real world scenario.
An analysis of the distribution of the length of mouth gestures can be seen in figure 2. It shows the box plot of the distribution. The length is given in number of frames, where all videos have been recorded at 50 frames per second. It can be seen that the majority of mouth gestures are relatively short. The mean is 24.9 frames and the 75th percentile 31 frames. Nevertheless the length can vary substantially. 5.1% of the mouth gestures have more than 60 frames and outliers reach up to 224 frames.
In our experiments we test a range of durations from 19 to 45 frames.

5. Experiments

5.1. General Preprocessing

We use OpenPose (Cao et al., 2018), (Simon et al., 2017) to detect facial landmarks on the face. These are used to transform the image so that the distance between the eyes is the same for all frames and all persons. We normalise the scale of the frames by the interocular distance and rotate them so that the axis between the eyes is horizontal. After alignment the ROI is extracted as described in 4.1.
The video clips are converted from RGB to grayscale, as previous tests showed no significant difference in the results. All frames are scaled to 96×96 pixels and normalised with the overall mean and standard deviation of the dataset. As the number of examples per class differ a lot, the dataset is balanced by over- or undersampling classes to the median of examples per category.

5.2. Training

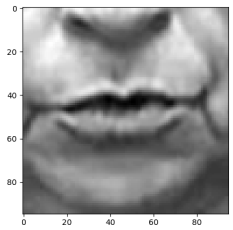
We use the pretrained model for the visual stream of (Petridis et al., 2018) and train the network end-to-end. The initial learningrate is set to $3 \cdot 10^{-4}$ in the ROI experiments and $3 \cdot 10^{-5}$ in the frame sampling and clip duration experiments, as the latter proved to be better in intermediate tests.
For data augmentation the data is cropped randomly to 88×88 pixels and randomly flipped horizontally during training.
As our dataset is small, we use 10-fold cross validation

mouth gesture	description	number of examples
MO04	Lips open and stretched, teeth together and visible. Like german 'sss' or 'pss'.	98
MO07/LR03	Lips round and open. Like german 'o' .	167
MO08	Mouth wide open. Like german 'a' .	113
LR01	Lips round and puckered, air streams out through small opening. Like german 'sch'.	52
LR02/LR10	Lips pursed.	420
LR05/CH01	Blow out air continuously through rounded lips, cheeks possibly puffed.	615
LC04/LC05	Lips closed and stretched strongly, lips possibly sucked in.	556
LC06	Lips closed, corners of mouth curved down, lips possibly sucked in.	340
TO01/TO04	Mouth open, tongue protrudes or dorsum pressed to front.	264
TE03	Mouth slightly open, upper teeth on lower lip, sudden release of air. Like german 'pf'.	216

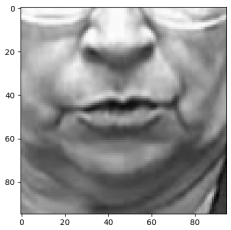
Table 1: Description of the ten mouth gesture classes used for classification and the number of annotated examples per class.

(a) original frame

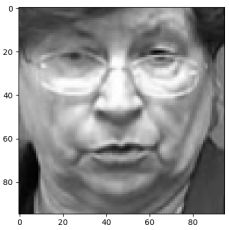
(b) pretraining ROI



(c) mouth ROI



(d) lower face ROI



(e) full face ROI

Figure 1: Example of an original frame (a) and our proposed ROIs (c)-(e). (b) shows the ROI used in (Petridis et al., 2018), with which our network was pretrained.

to make use of all data in the training- as well as in the validation set. This makes the results also more stable to statistical variations in the training procedure as all results are the combination of 10 individual training runs. Due to the dataset size we do not use a testset. All given results are the peak accuracy on the combined validation sets.

5.3. Experiments on ROI

Clips are cut to 29 frames using the up- and downsampling method described in 4.2.
When using a pretrained model one might achieve better results with inputs similar to the previous training material. Our network was pretrained using a ROI that is similar to our 'mouth only' ROI, see Figure 1. To ensure that we choose the best ROI for our dataset and not simply the one closest to the pretraining data, we run our experiment twice. The first time we use the pretrained model, the second time we train the network from scratch, to avoid influence from pretraining.

5.4. Experiments on Frame Sampling

For the experiments on the frame sampling methods we use the lower face ROI and cut the videos to 29 frames using either sampling method. We use the same sampling method for training and validation set.

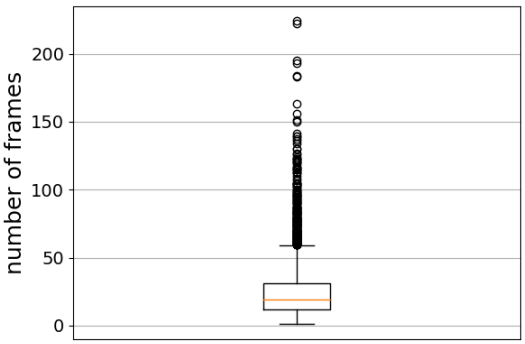


Figure 2: Box plot of the number of frames per mouth gesture for a frame rate of 50.

5.5. Experiments on Clip Duration

We use the' lower face' ROI and consider a number of frames from 19 to 45. For training, the videos are cut using the exact timing information given by the annotators. We cut the videos so that the midpoint matches with the midpoint of the true location of the mouth gesture. Therefore the clips may show more or less than the mouth gesture, but the mouth gesture is centred within the clip.
For the validation set we use two different options. One is to center the clip at the midpoint of the mouth gesture, as done with the training set. However, this information is not available in a real world scenario. Here we can only use the midpoint of the hand gesture as an approximation of the midpoint of the mouth gesture. This change may have an influence on the optimal number of frames, as more frames might be needed, to ensure that enough of the mouth gesture is included, if the timing of the mouth and hand gesture differ significantly. Therefore we validate the training results with both cuts.
Due to timing issues, the dataset was updated during the experiments. We therefore run the experiment with 29 frames twice, once on the old and once on the new version of the dataset to make the results comparable.
As the dataset used for pretraining consists of clips with a length of 29 frames, this might influence the results. We therefore rerun part of the experiment with the network trained from scratch.

6. Results

6.1. ROI

Table 2 shows the results for the three different ROIs with and without pretraining. For all ROIs the pretrained results are clearly better. In both the untrained and pretrained case the 'full face' ROI results in the lowest accuracy. So if there are helpful features on the upper part of the face, they are not strong enough to compensate the lower resolution of the images and the inclusion of unnecessary artefacts such as hair.
Without pretraining the 'mouth only' ROI gives better results than the 'lower face' ROI. With pretraining the 'lower face' ROI is better. This is a surprising result as the ROI used for pretraining is more similar to the 'mouth only' ROI. Therefore, the cause for the better results of the 'lower face' ROI in the pretrained case can not be, that the inputs are more similar to the pretraining inputs. Instead, the reason might be that a larger ROI is more complex to analyse. So the untrained network might fail to find the right features here and prefer more focused images, while the pretrained network already learned to find these features with the help of a much larger dataset and therefore benefits from the larger ROI with more features.
Therefore the 'lower face' ROI is preferable when using the pretrained network.

ROI	without pretraining	with pretraining
whole face	58.18	66.76
lower face	60.56	**70.60**
mouth only	**62.08**	68.93

Table 2: Peak accuracy for different ROIs.

6.2. Frame Sampling

The results for the frame sampling methods can be found in Table 3. The up- and downsampling method reaches a peak accuracy of 69.89 %, the continuous method 70.28 %. So the results for the continuous method are slightly better, but there is no significant difference. It seems, that it is at least equally important, that the clips are consecutive, to that the mouth gesture is cut exactly from start until end. The reason for that might be that the spatiotemporal convolution works best for consecutive frames. If several frames are doubled the layer can not extract any temporal information. If too many frames are deleted the facial movements might be too large and important frames might be skipped.

sampling method	accuracy
up-/downsampling	69.89
continuous	**70.28**

Table 3: Peak accuracy for different sampling methods.

6.3. Clip Duration

Table 4 shows the results for different clip durations, ranging from 19 to 45 frames with a frame rate of 50 frames per second. Here the pretrained model was used as starting point. As described in section 5.5, we used two different versions of the dataset and ran the experiment twice with 29 frames to make the results comparable. The change of dataset is indicated by the dashed line in the table.
When the videos are cut using the hand gesture position the accuracy decreases as expected. It is on average 2.7 percentage points lower. Apart from that, the results for both cuts are very similar.
Using less than 29 frames clearly worsens the results. For

29 to 39 frames there is no significant difference in the achieved accuracy. For 45 frames the accuracy drops for both cuts. We assume that for more than 39 frames the clips involve too much other material, that is not part of the mouth gesture.

Table 5 shows the classification accuracy for different numbers of frames, when the network is trained from scratch. Again, the results for the hand gesture cut are less accurate than for the mouth gesture cut, but apart from that the results are similar. However, when the network is trained from scratch shorter clips are clearly preferable. Here 19 frames achieve a better result than 29 or 39 frames. The reason might be that the pretrained network prefers 29 frames because that is what it was pretrained on. However, an argument against that is, that 35 and 39 frames achieve similar results in the pretrained case. Another, possibly additional, reason might be, that the untrained network prefers 19 frames because that is less data to process and features are easier to spot, which is not such a problem for the pretrained network. To find out which is the case here, it would be necessary to cut the videos of the LRW dataset to less frames, train the network with it and use this as a pretrained model for further experiments. However, this is beyond the scope of this work. Another possibility would be to create a model that is less biased to the number of frames in the clips, by performing variable length augmentation as described in (Martinez et al., 2020). In this case it might also be beneficial to use the mouth gesture data with its actual varying length. For the training data the exact length is known, for application on unlabelled data, it might be possible to estimate it by the hand gesture length, if available.

number of frames	mouth gesture cut	hand gesture cut
19	68.07	65.29
25	69.14	66.49
29	70.28	67.16
29	70.47	**68.04**
35	70.40	67.76
39	**70.78**	67.94
45	69.55	67.33

Table 4: Peak accuracy for different number of frames, using the pretrained model.

number of frames	mouth gesture cut	hand gesture cut
19	**58.03**	**55.82**
29	55.25	53.07
39	53.45	51.02

Table 5: Peak accuracy for different numbers of frames, training from scratch.

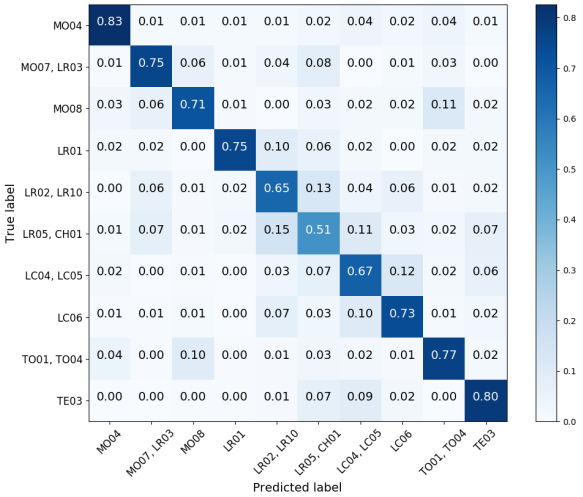


Figure 3: Confusion matrix for classification with continuous 29 frames centred at the hand gesture position.

6.4. Overall Results

Combining the results from all experiments, the best results are achieved with the pretrained model when using the 'lower face' ROI together with the continuous sampling method and a duration of 29 frames, which is equal to 0.58 seconds. For this setting we achieve an accuracy of 70.47% using the mouth gesture cut and 68.04% using the hand gesture cut, which would be used in a real world scenario. These results are comparable to human accuracy, which is between 66.47% and 79.13% for the given dataset.

Figure 3 shows the confusion matrix for the latter setting. It can be seen that the per class accuracy varies substantially from class to class, as some classes are well-defined while others overlap. For example, the round lips in class LR02/LR10 are similar to the lip shape when blowing air, as in LR05/CH01. The vibrating lip pattern of MO04 on the other hand, is unique and therefore easier to distinguish.

Interestingly, the number of examples per class in the dataset does not seem to have a high impact on the per class accuracy, as MO04 has the second least number of examples but the highest accuracy, while the three classes with most examples have the lowest accuracy.

7. Conclusion

In this work we compare different preprocessing options for mouth gesture classification from video.

The experiments on using different ROIs show that the best results can be achieved with a ROI that shows the lower half of the face. We compared two methods to format the videos to a fixed length: up- or downsampling the frames, so that the mouth gesture is shown exactly from start until end or using a time window of continuous frames centred to the middle of the mouth gesture duration. Both show similar results. We favour the continuous method, as it requires less information. Another important parameter is the duration of the videos, to make sure the relevant parts of the mouth gesture are included, but not too much additional material. We tested a range of 19 to 45 frames and showed that a

length of 29 frames is best, when using the pretrained network. When training from scratch, less frames are preferable.

Combining the results of all our experiments we achieve the highest accuracy when using the 'lower face' ROI and a duration of 29 continuous frames. If the clips are centred with the information of the exact mouth gesture we achieve an accuracy of 70.47%. If we use this information for training only and not for testing, as would be the real world scenario, the accuracy is 68.04%. In both cases we achieve results comparable to human accuracy, which lies in between 66.47% and 79.13%.

8. Acknowledgements

This publication has been produced in the context of the joint research funding of the German Federal Government and Federal States in the Academies' Programme, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies' Programme is coordinated by the Union of the German Academies of Sciences and Humanities.

9. Bibliographical References

Afouras, T., Chung, J. S., Senior, A., Vinyals, O., and Zisserman, A. (2018). Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*.

Brumm, M., Johnson, R., Hanke, T., Grigat, R.-R., and Wolfe, R. (2019). Use of avatar technology for automatic mouth gesture recognition. In *SignNonmanuals Workshop 2, Graz, Austria, 2019*.

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., and Sheikh, Y. (2018). OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*.

Chung, J. S. and Zisserman, A. (2016). Lip reading in the wild. In *Asian Conference on Computer Vision*.

Chung, J. S., Senior, A., Vinyals, O., and Zisserman, A. (2017). Lip reading sentences in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Koller, O., Ney, H., and Bowden, R. (2014). Read my lips: Continuous signer independent weakly supervised viseme recognition. In *European Conference on Computer Vision*, pages 281–296. Springer.

Koller, O., Ney, H., and Bowden, R. (2015). Deep learning of mouth shapes for sign language. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 85–91.

Luzardo, M., Viitaniemi, V., Karppa, M., Laaksonen, J., and Jantunen, T. (2014). Estimating head pose and state of facial elements for sign language video. In *6th Workshop on the Representation and Processing of Sign Languages: Beyond the Manual Channel, Language Resources and Evaluation Conference*.

Martinez, B., Ma, P., Petridis, S., and Pantic, M. (2020). Lipreading using temporal convolutional networks. *arXiv preprint arXiv:2001.08702*.

Neidle, C., Liu, J., Liu, B., Peng, X., Vogler, C., and Metaxas, D. (2014). Computer-based tracking, analysis, and visualization of linguistically significant nonmanual events in american sign language (ASL). In *6th Workshop on the Representation and Processing of Sign Languages: Beyond the Manual Channel, Language Resources and Evaluation Conference*.

Petridis, S., Stafylakis, T., Ma, P., Cai, F., Tzimiropoulos, G., and Pantic, M. (2018). End-to-end audiovisual speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018*, pages 6548–6552. IEEE.

Simon, T., Joo, H., Matthews, I., and Sheikh, Y. (2017). Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1145–1153.

Von Agris, U., Knorr, M., and Kraiss, K.-F. (2008). The significance of facial features for automatic sign language recognition. In *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 1–6. IEEE.

Wilson, N., Brumm, M., and Grigat, R.-R. (2019). Classification of mouth gestures in german sign language using 3d convolutional neural networks. *International Conference on Pattern Recognition Systems ICPRS 2019*.

Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 33–38
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

PE2LGP Animator: A Tool to Animate a Portuguese Sign Language Avatar

Pedro Bertrand Cabral, Matilde Gonçalves, Ruben dos Santos, Hugo Nicolau, Luisa Coheur
Instituto Superior Técnico, Universidade de Lisboa/INESC-ID
{pedro.b.cabral, matilde.do.carmo.lages.goncalves, luisa.coheur, hugo.nicolau}@tecnico.ulisboa.pt, soltex@me.com

Abstract

Software for the production of sign languages is much less common than for spoken languages. Such software usually relies on 3D humanoid avatars to produce signs which, inevitably, necessitates the use of animation. One barrier to the use of popular animation tools is their complexity and steep learning curve, which can be hard to master for inexperienced users. Here, we present PE2LGP, an authoring system that features a 3D avatar that signs Portuguese Sign Language. Our Animator is designed specifically to craft sign language animations using a key frame method, and is meant to be easy to use and learn to users without animation skills. We conducted a preliminary evaluation of the Animator, where we animated seven Portuguese Sign Language sentences and asked four sign language users to evaluate their quality. This evaluation revealed that the system, in spite of its simplicity, is indeed capable of producing comprehensible messages.

Keywords: Portuguese sign language, avatar, animation, accessibility

1. Introduction

PE2LGP is a project to digitalise *Portuguese Sign Language* (shortened to *LGP – Língua Gestual Portuguesa*), the primary language of the Deaf community in Portugal, through a 3D avatar capable of communicating in it.

Though a living language used by thousands of people, LGP is still largely understudied, with both an absence of linguistic research on it (compared to widely used spoken languages) and a lack of tools and resources for its computational processing. We aim to provide one such resource through this project, in this case through the Animator, a tool that allows users without technical knowledge or animation expertise to create animations of LGP signs for our avatar to use through simple frame-by-frame posing. As part of the larger effort to improve digital support for LGP, the Animator could be used for chatbots, virtual assistants, dictionaries, or, as we have done in PE2LGP, automatic translators.

The motivation for this study was to not only expose this project to the community, but to gain greater intuition of the tool's current performance. We thus present in this paper a description of our Animator and an overview of the role it plays in the greater scope of LGP and sign language research, along with a preliminary study of its capabilities, where, using our tool, we animated seven LGP sentences and asked four users of the language to interpret them and give an appreciation of their quality.

2. State of the Art

Sign languages are visuospatial languages, i.e., the communication is performed using signs produced at determined locations in the three-dimensional space or on the body. Because of the deafs' linguistic isolation, they must sometimes resort to human interpreters, who are not always available, so alternative translation systems are useful. Sign languages have no widely-used written forms (Kaur and Kumar, 2014), so such systems require the representation of a human body to produce messages, such as videos of human signers or 3D avatars.

However, videos of human signers have serious limitations when stringing together signs to compose sentences (Huenerfauth and Hanson, 2009). First, because each sign would have to be recorded in its own video, it would not be possible to have smooth movement from one sign to the next: the signer would have to return to a neutral position and there would likely always be seams. Furthermore, the videos would have to be recorded with the same signer under similar conditions, to avoid abrupt visual changes. Finally, this approach would not allow different signs to be combined: in LGP, for example, a facial expression may be combined with other signs to mark a sentence as interrogative. An avatar-based solution, on the other hand, is more flexible and sidesteps all the problems pointed out in the previous approach. The main concern with an avatar is the work required to create animations that can be combined to produce clear, natural-looking signing.

Several methods exist to virtually recreate of sign language gestures with 3D avatars. They can be clustered into three types: hand-crafted, motion capture and synthesis from a sign notation system (Gerlach et al., 2016).

Motion capture is done by recording signs made by a human using video cameras or other types of sensors and later mapping the human actor's motions onto an avatar. The more detailed the desired result, the costlier and more complex the technology and expertise required. Motion capture frequently requires calibration and must usually be used alongside hand-crafted animation, because the resulting performance often has to be fine-tuned, especially when using cheaper solutions, such as Kinect and Leap Motion (Gerlach et al., 2016). Caroline Guardino and Ching-Hua Chuan attained better results using Leap motion than Kinect and Cyblerglove, used in other studies for recognising sign language (Guardino et al., 2014).

The second type of animation consists in creating a system capable of automatically interpreting a phonetic sign language writing system, like HamNoSys (Hamburg Sign Language Notation System) to animate a signing avatar (Zwitserlood et al., ; Elliott et al., 2004). This writing system gives us detailed information about the hands elements

and other human movements that compose a sign (Hanke, 2004), but not secondary movement (unlike motion capture).

Lastly, hand-crafted animation is the oldest of these techniques, widely used, and known to give good results, but also requires intensive work, as someone must manually pose the avatar and adjust the animation until the result is satisfactory. The more realistic and detailed the animations, the more time, effort, expertise and technological sophistication are necessary. Blender[1] and Unity[2] are widely used general-purpose 3D computer graphics software capable of animating avatars. Blender in particular has a vast feature set, but also a steep learning curve (Cano, 2011). In contrast, our Animator is designed specifically for animating humanoid characters, which allows us to restrain its complexity.

3. PE2LGP Description

3.1. Overview

PE2LGP was originally created by Inês Almeida (2014) and further built upon by Ruben Santos (2016) in their master's theses. The project is part of *Corpus & Avatar da Língua Gestual Portuguesa*, a joint effort of researchers at Instituto Superior Técnico and Universidade Católica Portuguesa to create not only an avatar capable of signing LGP, but also the first LGP *corpus*, complete with video, translation, gloss and syntactic annotation. This interdisciplinary approach of linguistics and computer science allows greater cooperation between two otherwise separate projects, with the corpus being used for applications such as machine translation and animation synthesis through HamNoSys.

PE2LGP currently has 5 components, all of which feature our 3D avatar, *Catarina*, as a centrepiece. These are the Translator, the Animator, the Hand Pose Editor, the Kinect Recorder and the Animation Viewer. The Translator component receives a sentence in Portuguese as input, which will then be automatically translated to LGP and signed by Catarina. The Animator, being the focus of this paper, is described in detail in Section 3.2., and it allows the user to create new animations using *forward kinematics* (manipulating each of the avatar's joints individually). The Hand Pose Editor (created in the time between the execution of this study and its final revision) allows users to create and modify hand poses, which are then used in the Animator. The Kinect Recorder's purpose is to create new animations through rudimentary motion capture using a Kinect device. The Animation Viewer is a simple menu which allows the user to view and delete existing animations.

The Animator plays a crucial role in the project as the principal tool for creating signs for our avatar, with the ultimate goal of creating an animation database to be used by the Translator component or any other future component that requires it (such as dictionaries, messaging systems, chat bots or games).

3.2. Animator

Our Animator component makes use of key frames, premade hand poses, and forward kinematics. You can see a screenshot of the animator's interface in Figure 2.

Key frames are the principal poses the avatar will assume throughout the time span of an animation (at a pace of one key frame per second). The user must define a key pose for each key frame one by one and, when the animation is played back, the avatar will not only assume the correct key pose for each key frame, but also automatically interpolate between key poses to generate all the in-between frames (Figure 1). This key frame method allows the user to focus on the essential moments of the sign and only pay attention to the intermediate moments when the situation requires it.

Forward kinematics is used by the system to allow the user to manipulate the avatar into the desired key poses by rotating the avatar's joints so, for example, rotating an elbow will move the forearm but leave the upper arm in place. The joints available for the user to manipulate are the neck, waist, shoulders, elbows, and wrists, which can all be rotated in 3 axes.

By design, the editor does not permit manipulating the joints on the avatar's hands and fingers directly (only its wrists). Instead, the user must choose from a selection of pre-made hand poses, which may be changed each frame and chosen independently for the right and left hand (Figure 3). When this experiment was conducted, these hand poses were limited to a selection of older animations created at an earlier stage of the project but, at the time of writing, a new component called the Hand Pose Editor has been fully implemented, enabling users to create and modify hand poses at will.

The project does not yet support facial expressions, though this feature is a current priority.

4. Evaluation Methodology

Seven simple LGP sentences (Table 1) were created and animated by two engineers with basic LGP training working on the project, using the corpus (Section 3.1.) and two online dictionaries as reference[3,4]. Most signs for these sentences were newly-created using the Animator, but some were already present on the platform's database, such as the letter signs used for finger-spelling[5]. Each new sign took between 10 and 60 minutes to animate, depending on its complexity and the desired quality and detail.

To string together signs to form sentences, Unity's native animation features were used, specifically its *animation controller* system, which employs state machines to allow different animations to be played back and mixed. Without this mechanism, the transitions between signs would have been less smooth, because the avatar would have been forced to return to a neutral position (standing upright, with arms at its sides) between every sign. Using this system, however, is not effortless, as it requires calibration dependent on which two animations the transition involves. Note the distinction between: interpolation between key frames

[1] www.blender.org

[2] unity.com

[3] www.spreadthesign.com/pt.pt

[4] www.infopedia.pt/dicionarios/lingua-gestual

[5] Finger-spelling consists of signing individual letters of the alphabet to spell out words, usually proper names. In Table 1 you can see that the proper name "Júlio" was finger-spelled as J-U-L-I-O.

Figure 1: In the Animator, the user defines only key frames A and B, and the system interpolates between those two key frames to create the in-between frames, as the image illustrates. Note that this is a simplified view: in reality, there would normally be 58 in-between frames, not 4.

Figure 2: Screenshot of the Animator. Notice the red hoop around the avatar's waist, indicating the selected joint and how it will rotate. On the right, you can see the X, Y, and Z buttons, which control which axis is to be rotated. On the left, there are various controls, such as for creating new frames, switching hand poses, and previewing the animation.

(which occurs *within an animation*) and transitions *between two animations*. The former occurs in the Animator, when the sign is created, while the latter is done in Unity's animation controller.

The grammatical correctness of the 7 sentences was then verified by a hearing linguist with intermediate mastery of LGP and a trained LGP interpreter (both native Portuguese speakers). This validation yielded several recommendations, which were then implemented within the boundaries of the Animator before being committed to the next phase.

Finally, videos (which are available at this footnote URL[6]) of the avatar signing the sentences were sent to 4 evaluators, who answered an online form about the avatar's quality without being told the sentences' meaning or which signs were being produced. The evaluators described their proficiency as "high", "moderately proficient", "medium" and "professional". The form consisted of a set of questions which were repeated for each sentence: the evaluator would view the video of the avatar signing and try to determine the meaning of the sentence. Whether or not the evaluators were able to discern a sentence's meaning was the main factor in measuring how intelligible it was. Next, the evaluator would answer how many times they had to watch the video and rate several aspects of the sentence's quality: speed, overall quality, intelligibility, naturalness, grammar, hand configuration, hand orientation, hand location, and hand movement. These aspects were rated on a scale of 1 to 5 with 1 being "Bad" and 5 being "Good" (speed was the exception, where 1 meant "Too slow" and 5 meant "Too fast".

[6] tinyurl.com/PlaylistAvatarLGP

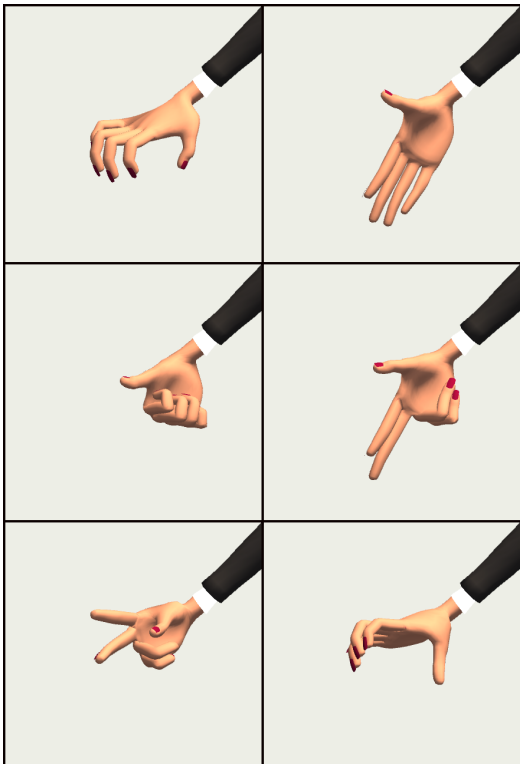


Figure 3: Examples of pre-made hand poses, which can be selected in the editor.

5. Results

The full quantitative responses to our survey are available in Table 3 and Table 2.

Translation	LGP (gloss)
Bom dia (Good day)	BOM DIA GOOD DAY
A rainha é soberana (The queen is sovereign)	MULHER REI IMPORTANTE MUITO CHEFE WOMAN KING IMPORTANT VERY BOSS
O rapaz chegou à cozinha (The boy arrived at the kitchen)	RAPAZ COZINHA CHEGAR BOY KITCHEN ARRIVE
O meu nome é Júlio (My name is Julio)	NOME MEU J-U-L-I-O NAME MINE J-U-L-I-O
O rei pensou e teve uma ideia (The king thought and had an idea)	REI PENSAR TER-IDEIA KING THINK HAVE-IDEA
O meu irmão come peixe (My brother eats fish)	IRMÃO MEU PEIXE COMER BROTHER MINE FISH KING
Ele dança bem (He dances well)	ELE DANÇAR BEM HE DANCE WELL

Table 1: Sentences used in the evaluation.

Naturalness was the lowest-rated aspect, with an average of 1.7. This was to be expected, as the avatar's movements are too machine-like to appear human. The main requirements to improve naturalness would be facial expressions (both grammatical and non-grammatical), configurable key frame interpolation, automatic secondary movement, lower-body movement.

The highest-rated aspect was *Speed*, with 18 out of 28 perfect classifications. Part of the reason for this may be that speed problems are localised to particular signs or transitions. Another factor may be the initial waiting time within the videos, before the avatar begins signing. This waiting time was unintentional and not consistent across the videos (ranging from 4 seconds to less than a second), and we suspect it may create the illusion of a slower animation.

We consider our most successful sentence to be ELE DANÇAR BEM, which was almost perfectly understood by all evaluators with a single viewing of the video and consistently outperformed other sentences across all categories. This sentence is interesting, as it is the only one in this experiment to include non-manual movement (waist and neck motion), and we believe that is why it scored higher in Naturalness than any other sentence.

We consider the least successful sentence to be the second, MULHER REI IMPORTANTE MUITO CHEFE, which the evaluators viewed more times than usual and had trouble understanding (one could not name any signs correctly). One suspected cause for this difficulty is the absence of context.

In the categories of hand configuration, hand orientation, hand location and hand movement, the responses often corresponded to our expectations, where sentences with higher quality signs received better scores. In a few cases, the responses (both quantitative and open-ended) led us to discover more subtle improvements that could be made to the signs (such as bringing the hand closer to the chin in the sign BOM).

In some sentences, we detected that unruly transitions between animations made signs less clear, or at least less natural, by making the avatar's motion too quick or anatomically impossible. A number of problems with individual signs were also detected, most often imprecise hand poses.

6. Conclusion

In this paper, we presented our Animator tool which, as a part the PE2LGP project, aims to be an accessible means to animate signs to be performed by an avatar. We also performed a preliminary evaluation of this system using Portuguese Sign Language sentences which, although too small to yield statistically significant results, provides valuable insight into the current capabilities, limitations and future potential, while demonstrating that the platform is indeed capable of producing comprehensible LGP, which is a positive result, given its simplicity and the complexity of synthesising natural languages.

In the future, we would like to improve the Animator with both quality-of-life features, to make editing animations more comfortable and efficient, and, more importantly, improvements to enhance the tool's capacity, such as custom hand posing, facial expressions, and control over interpolation. For further research, it would be interesting to formally study the Animator's ease of use through user tests, as accessibility is one of its main design goals. Ultimately, this usability should enable the development of a thorough database of animations by using the Animator and its companion components to swiftly bring the first-hand knowledge of LGP users into the platform, to be used as a resource in this and other Portuguese Sign Language projects.

7. Acknowledgements

We owe our thanks to our four evaluators, who took the time to help in this study, and to Mara Moita and Neide Gonçalves, who helped ensure our LGP sentences were correct.

This work was supported by national funds through FCT, Fundação para a Ciência e a Tecnologia, under project UIDB/50021/2020 and PTDC/LLT-LIN/29887/2017.

8. Bibliographical References

Almeida, I. R. (2014). Exploring challenges in avatar-based translation from european portuguese to portuguese sign language. Master's thesis, Instituto Superior Técnico, October.

Category	Evaluator	S1	S2	S3	S4	S5	S6	S7
Number of Views	*A*	2	>1	3-4	3-4	4	4-5	1
	B	1	7	2	3	2	1	1
	C	1	>1	>1	>1	>1	>1	1
	D	2	5-6	5	1	3	1	1
Speed	*A*	2	1	1	2	1	2	3
	B	3	1	3	2	2	3	4
	C	4	NA	2	3	2	3	4
	D	3	NA	1	2	2	3	5

Table 2: These were the responses of each evaluator to the questions "How many times did you need to watch the video?" and "How adequate was the sentences's speed?". Unlike the results presented in Table 3, the first question was open-ended, while the second was on a scale of 1 to 5, with 1 labeled "Too slow" and 5 labeled "Too fast".

Category	Evaluator	S1	S2	S3	S4	S5	S6	S7
Overall Quality	*A*	2	1	1	2	1	2	3
	B	3	1	3	2	2	3	4
	C	4	NA	2	3	2	3	4
	D	3	NA	1	2	2	3	5
Inteligibility	*A*	2	1	2	1	1	2	3
	B	4	1	3	4	3	4	5
	C	4	2	2	2	2	3	4
	D	4	NA	1	3	2	3	5
Naturalness	*A*	1	1	1	1	1	1	2
	B	1	1	1	2	2	1	2
	C	2	2	2	2	1	2	2
	D	2	NA	1	2	3	3	3
Grammar	*A*	3	1	2	3	NA	2	3
	B	5	NA	4	5	4	4	4
	C	4	NA	3	4	NA	4	4
	D	5	NA	5	5	3	5	5
Hand Configuration	*A*	3	2	1	1	1	1	3
	B	5	2	2	1	3	1	5
	C	4	NA	2	1	2	2	NA
	D	5	NA	1	4	3	4	5
Hand Orientation	*A*	3	1	1	2	1	2	3
	B	4	2	5	3	2	3	5
	C	4	NA	NA	NA	2	2	NA
	D	5	NA	2	5	4	5	5
Hand Location	*A*	1	1	2	2	1	1	3
	B	4	4	5	4	2	3	5
	C	4	NA	NA	1	3	2	3
	D	4	NA	2	5	4	4	3
Hand Movement	*A*	1	1	1	1	1	1	3
	B	3	4	1	2	2	1	5
	C	4	NA	2	2	2	1	3
	D	4	NA	2	4	3	4	5

Table 3: These were the responses of each evaluator for each category and sentence (*S1*, *S2*, etc), on a scale of 1 to 5. In the survey, the response *NA* was written "I don't know", while 1 was labeled "Bad" and 5 was labeled "Good".

Cano, J. L. C. (2011). The cambrian explosion of popular 3d printing. *IJIMAI*, 1(4):30–32.

Elliott, R., Glauert, J. R. W., Kennaway, J. R., and Marshall, I. (2004). The development of language processing support for the ViSiCAST project. pages 101–108.

Gerlach, J., Strasly, I., Ebling, S., Rayner, M., Bouillon, P., and Tsourakis, N. (2016). An Open Web Platform for Rule-Based Speech-to-Sign Translation. (August):162–168.

Guardino, C., Chuan, C.-H., and Regina, E. (2014). American sign language recognition using leap motion sensor. 12.

Hanke, T. (2004). HamNoSys-Representing sign language data in language resources and language processing contexts. *Proceedings of the Workshop on Representation and Processing of Sign Language, Workshop to the forth*

International Conference on Language Resources and Evaluation (LREC'04), pages 1–6.

Huenerfauth, M. and Hanson, V. L. (2009). Sign language in the interface: Access for deaf signers. *The Universal Access Handbook*, pages 38–1–38–18.

Kaur, R. and Kumar, P. (2014). Hamnosys generation system for sign language. In *2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 2727–2734, Sep.

Santos, R. E. R. (2016). Pe2lgp: Do texto à língua gestual (e vice-versa). Master's thesis, Instituto Superior Técnico, October.

Zwitserlood, I., Verlinden, M., Ros, J., Schoot, S. V. D., and Netherlands, T.). SYNTHETIC SIGNING FOR THE DEAF: eSIGN.

Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 39–44
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

Translating an Aesop's Fable to Filipino Sign Language through 3D Animation

Mark Cueto, Winnie He, Rei Untiveros, Josh Zuñiga, Joanna Pauline Rivera
De La Salle University - Manila
2401 Taft Avenue, Malate, Manila, Philippines 1004
{winnie_he, joanna.rivera}@dlsu.edu.ph

Abstract

According to the National Statistics Office (2003) in the 2000 Population Census, the deaf community in the Philippines numbered to about 121,000 deaf and hard of hearing Filipinos. Deaf and hard of hearing Filipinos in these communities use the Filipino Sign Language (FSL) as the main method of manual communication. Deaf and hard of hearing children experience difficulty in developing reading and writing skills through traditional methods of teaching used primarily for hearing children. This study aims to translate an Aesop's fable to Filipino Sign Language with the use of 3D animation resulting to a video output. The video created contains a 3D animated avatar performing the sign translations to FSL (mainly focusing on hand gestures which includes hand shape, palm orientation, location, and movement) on screen beside their English text equivalent and related images. The final output was then evaluated by FSL deaf signers. Evaluation results showed that the final output can potentially be used as a learning material. In order to make it more effective as a learning material, it is very important to consider the animation's appearance, speed, naturalness, and accuracy. In this paper, the common action units were also listed for easier construction of animations of the signs.

Keywords: Filipino Sign Language, text to sign language, 3D animation

1. Introduction

In the 2000 Population Consensus, the National Statistics Office (2003) indicated in their report that there are 121,000 deaf and hard of hearing Filipinos. With the use of sign language, it became a possibility for them to communicate and, in a way, express themselves to other people. According to Mendoza (2018), these deaf and hearing Filipinos use various languages to communicate such as American Sign Language (ASL), Signing Exact English (SEE), and Filipino Sign Language (FSL). Among these methods, FSL is the most widely used at a percentage of 70% of Filipino signers using it as their main sign language (Imperial, 2015).

A wide range of deaf and hard of hearing individuals have difficulties in reading. This is because written text follows different grammatical structures from that of FSL, causing confusion especially to deaf and hard of hearing children who are only starting to read (Flores, 2012). As stated by Imperial (2015), signers first learn FSL for communication, a visual-spatial language different from an auditory-vocal language like Filipino (and its written transcripts).

Deaf and hard of hearing children experience difficulty in developing reading and writing skills through traditional methods of teaching used primarily for hearing children (Mich et al., 2013). A reason for this is because they cannot hear and distinguish the phonemic sound system or what the spoken language sounds like (Magee, 2014). Another reason is the method of teaching that is given to them, in which the learning method teaches them how to understand single words and single sentences instead of learning the full text. Thus, specialized methods catered specifically to their needs must be utilized to aid in the development of reading skills (Mich et al., 2013).

In order to help with the development of vocabulary and reading skills in young signers, supervision from an adult signer can provide a significant improvement. According to Huff (2012), older signers who can provide translations from the text into FSL could aid in the child's understanding of the text and in the long run, would help in the development of reading skills.

An example of text that children enjoys are stories. These are one of the most prevalent sources of knowledge. They help children learn and make sense of the world (Lonneker and Meister, 2005). Fables are stories that teach important life lessons using animals as characters (Ang et.al, 2010). By using animals instead of human characters, readers would less likely be biased when reading the story. It allows readers to go through the story without thinking of comparing themselves with the character. A famous set of fables are Aesop's fables, which is used for this study.

To teach the meaning of the story text, it must be translated to sign language first. An example of application that has already been developed for such purpose is MMSSign produced by Jemni, Ghoul, Yahia, and Boulares (2007) to aid in the use of cellular phones by those of the Deaf Community. It turns messages into a 3D animated video with the corresponding sign language, removing the need to read the contents of the message. This serves as a better system of communication for the signers.

In the Philippines, however, there is very minimal research that focuses on such technology that translates English text to FSL. According to Martinez and Cabalfin (2008), there are only few researchers that conduct studies for the Filipino deaf community.

The use of 3D animation in the delivering of translated passages (whether from spoken or written sources) is very widespread. Along with the use of videos depicting real people performing signs, it is the most commonly used in translation systems. Thus, 3D animation was utilized for the translation in this study.

This study aims to translate an Aesop's fable to FSL with the use of a 3D animation software resulting to an animated video output. Among the 5 components of FSL, which includes the hand shape, location, palm orientation, movement, and non-manual signals, the animation mainly focused on the hand gestures (includes hand shape, location, palm orientation, and movement).

The contributions of this paper are:

- a 3D animated version of a translated Aesop's fable to FSL which can serve as a learning material,
- a proposed layout for an FSL learning material that allows new signers to learn FSL signs by

associating the images with the FSL signs and to slowly learn the English text equivalent at the same time, and

- a list of the common hand gestures in FSL which can help in constructing 3D animated FSL signs.

2. Review of Related Literature

Multiple application and software have been developed in order to ease the communication between signers and non-signers. Addressing the gap in communication between the two communities are made easier through the use of technology. Through these technological methods, the need for an interpreter decreases as the translating power of devices increases. These applications can vary from translating text to Sign Language, or speech to Sign Language, or Sign Language to text. In the process of translating oral or written text into sign language, the transcript must first be transformed into the grammatical structure used in sign language before finding gestures to form the same message in sign language. These gestures are obtained from a database and most commonly presented in one of two ways: (1) a pre-recorded video of a person performing that sign stitched together to form a video, or; (2) a 3D animated model. For the reverse (Sign Language to Text), the use of motion capture devices is necessary to read the gesture performed before being translated into text.

2.1 Translation system by Halawani

In the study by Halawani (2008), the process of translation systems would be primarily divided into two parts, the conversion and the translation. Conversion refers to the process in which the written syntax of a language is dissected and arranged into the syntax of the signs, and translation changes the words into animated rendering of the sign language equivalent which is shown to the user on the screen of their device. The conversion process is a common challenge in producing translation systems, as most sign languages are not yet thoroughly studied in the area of its syntax and grammatical structure.

2.2 LODE-2

In the study of Mich, Pianta, and Mana (2013), they have developed a tool called LODE-2 which is an improved version of LODE-1. It is a system with interactive stories and exercises for deaf children. The system has dynamic feedback for improving the reading comprehension skills of deaf children. After reading a whole story, the children are instructed to solve some three exercises for assessment : a global comprehension of the story, a comprehension of local-temporal relations, and a comprehension based on pure text. They have concluded that simplified stories with illustrations is the most effective way in teaching and aiding deaf children's reading comprehension.

2.3 ATLASLang MTS 1

Brour and Benabbou (2018) created a machine translation system Arabic Sign Language (ArSL) which allows the input of Arabic text by a non-deaf user to be converted into ArSL and displayed using GIF Images. They developed ATLASLang MTS 1, an example-based and rule-based Interlingua approach system. The example-based is used when the given sentence exists in the database, otherwise, it uses rule-based Interlingua. For the animation, they translated sentences using a database of 200 words that are taken from the Moroccan Dictionary. However, if the given word is a proper noun or does not exist in the database, it will spell it out letter by letter (finger spelling). The researchers did an experiment using the ATLASLang MTS 1. In the experiment they conducted, not all sentences given were accurately translated. There were also cases that no results were given. Specific results were not shown. After their experimentation, the researchers concluded that to improve the current version of the system, they must expand their database and implement more rules. They also recommend to use 3D Human Avatar instead of GIF Images.

2.4 VGuido (eSIGN 3D Avatar)

Another translation system was that of San-Segundo et al. (2011) which translates spoken Spanish into Spanish Sign Language. According to San-Segundo et al. (2011), at least 92% of the Spanish deaf community have a hard time comprehending and conveying themselves in Spanish. Verb conjugations, abstract concept explanations, and gender concordance are just some of the problems that the Spanish deaf community is having. Another problem is that the accepted Spanish Sign Language is not spread well enough to other people resulting to a communication barrier between the Spanish signers and non-signers.

The objective of the study is to introduce the first system of translating Spanish speech into the Spanish sign language in assisting a deaf person in a kind of service like the renewing process of a driver's license. The system is composed of different parts for the system to work. The speech recognizer that converts the speech into a sequence of words, the natural language translator that translates the sequence of words into Spanish sign language, and the 3D avatar animation module that shows the sign language on the screen (San-Segundo et al., 2011).

In the system made by San-Segundo et al. (2011), three technological proposals for the natural language translator were utilized: (1) example-based strategy; (2) statistical translation, and; (3) rule-based translation. From a 0-5 scale, the user is asked if the signs are correct, if they understand the sequence of signs, if the signing is natural, and if they would use the system instead of a human. In the test, the system executed the task very well in speech recognition, only having a 4.8%-word error rate, and in language translation, only having an 8.9% sign error rate. Although the system executed the task very well, the people who used the system did not rate it with a very good score in the questionnaires with reasons being the avatar signs unnaturally and there were discrepancies. The researchers concluded that improvement is needed in the system, especially in the area of the 3D animation.

3. Methodology

Based on the related literature for sign language translation from text to sign, 3D animation is the most commonly used method of visualizing the translated passages. First, the 3D animation tool was selected. Second, the reference video for the signs was annotated. Third, the avatar was animated in accordance to the reference video. Fourth, the video elements (i.e. video of avatar signing, text equivalent, and other supplements) was compiled to one video.

3.1 Selection of 3D Animation Tool

A selection of 3D animation software was chosen and was tested and evaluated based on criteria needed for the research output. The criteria were based on the presence of features that were needed in making the animation which are detailed in Table 1.

SmartBody (Shapiro, 2015) lacked the drag and drop capabilities that could help ease the animation process. The complicated interface and the lack of avatar customization also made SmartBody unlikely to be chosen as the animation software for the project.

Alice (Alice, 2017) has a built-in character customization screen that can change specific parts of the avatar. The avatar included with the software is also pre-rigged, although, for example, the pinky finger in both hands does not have their own rigs and only moves based on the movement of rigs that are near or adjacent. This creates inaccuracies in the performed signs.

Blender (Blender Foundation, n.d.) does not provide a built-in character model but has an online community wherein royalty-free, rigged models are available for download. The animation capabilities of the software were also vast as it can animate up until the fingertips which is very important for FSL.

Through the testing process, Blender was chosen as the 3D animation software that will be utilized.

Tool	Ease Of Use	Simple Interface	Avatar Customization	Detailed Animation
Smart Body	No	No	Yes	Yes
Alice	Yes	Yes	Yes	No
Blender	**Yes**	**No**	**Yes**	**Yes**

Table 1: Features in the tested 3D animation software

3.2 Annotation of the Reference Video

The reference video used for this research is a publicly available video containing the FSL translation of the chosen story, "The Tortoise and the Hare" (Landayan, 2017). The video was annotated using EUDICO Linguistic Annotator (E-LAN) (Tacchetti, 2018).

There were three types of annotations made which are the common hand action units, glosses (i.e. words or phrases), and English sentences.

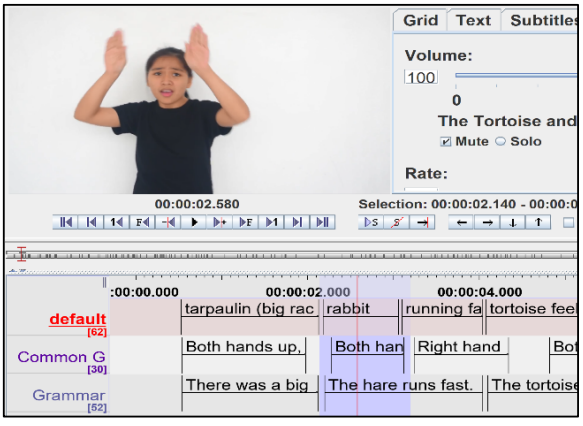


Figure 1: ELAN Annotation Interface.

3.2.1 Common Action Units

The hand action units were described in the annotation. Then, the reference video was analyzed to enumerate the common hand action units, and the unique hand action units. See Figure 1. For the word 'Rabbit', the person in the video raised both of her hands till her mid-body and her palm was facing outwards. Thus, what was entered in the tab was, "Both hands raised to mid-body, palms face outward". The list of the common hand gestures derived from the story is shown in Table 2. These may be used in animating a different story.

Hand Shape	Location	Palm Orientation	Movement
Both hands, open-A shape	Chest	Downward	Pointing to self
Both hands, X shape	Eyes	Inward	Circular motion
Both hands, 5 shape	Eyes	Inward	Fingers fluttering
Both hands, open-B shape	Chest	Center Line	Clapping
Both hands, S shape	Chest	Downward	Making an arc shaped motion forward
Both hands, open-B shape	Temples	Outward	
Right hand, 1 shape	Chest	Inward	Pointing to self/other
Both hands, L shape	Chest	Right hand, inward and Left hand, outward	Shifts to open-G shape
Right hand, open-A shape	Chest	Outward	Thumbs down
Right hand, B shape	Stomach	Inward/ Outward	Handshake
Right hand, open-B shape	Mouth	Inward	
Both hands, bent-5-claw shape	Chest	Upward	
Both hands, R shape	Forehead	Outward	Slight shake
Both hands, 5 shape	Arm sides	Center Line	Circular motion backwards
Both hands, open-B shape	Chest	Downward	Crawling
Both hands, S shape	Chest	Center Line	Circular Motion
Both hands, 1 shape	Chest	Center Line	
Left hand's thumb and index finger pinch together	Cheeks	Outward	Straight line motion outwards
Right hand, open-A shape and Left hand open-B shape resting at back of right hand	Chest	Right hand, center line and Left hand, downward	Slight shake for right thumb

Both hands, 5 shape	Right hand, nose and Left hand, chest	Right hand, center line and Left hand, downward	Slight shake for right hand
Both hands, 5 shape	Back of head	Outward	
Left hand, 1 shape	Chest	Outward	
Left hand, 5 shape	Face	Inward	Shifts to flat-O shape
Both hands, S shape	Chest	Downward	
Both hands, bent-5-claw shape	Neck	Inward	Fingers fluttering
Right hand, 4 shape	Face	Inward	
Both hands, C shape	Above head/ Chest	Center line	
Both hands, 5 shape	Chest	Inward	Brushing against body

Table 2: List of Common Gestures (left and right hand based on the person in the video)

3.2.2 Glosses

The gloss labels (i.e. words or phrases) were annotated by an FSL deaf signer for matching the performed signs to its corresponding English text as they are more knowledgeable. The FSL deaf signer is 22 years old with almost 5 years of experience in signing FSL.

However, the researchers searched for another annotator because the previous annotations made were mostly phrases or sentences which are made up of many different signs. This makes it difficult for the researchers to annotate its exact word or phrases. The other annotator is 25 years old and comes from a family of FSL deaf signers.

The annotation was done by first, selecting a (group of) sign performed in the video and entering the corresponding English text in the tab. For instance, since the sign performed in the selection corresponds to the word 'rabbit', the word 'rabbit' was entered in the tab (as shown in Figure 1). This step is important because not all English words have corresponding signs in FSL, as FSL has its own grammatical structure that focuses only on the main points of a sentence.

3.2.3 English sentences

After annotating the video to match the corresponding signs to its English equivalent, the reconstruction of words, phrases, and sentences was done as well to follow the English grammar. This was done by the researchers. For instance, in Figure 1, the words in the tab are "rabbit" and "running fast". Thus, following the English grammar, it was reconstructed to "The hare runs fast". This is an important step because FSL and English have a very different grammatical structure.

3.3 Animation of the Video

Using the annotated video as reference, the avatar was animated using Blender. The avatar used was a default model that Blender offers. The animation focused on the hand gestures only.

The animation is separated by, searching for common hand action units first. The researchers analyzed the video and categorized the common action units into hand shape, location, palm orientation, and movement. See Table 1 for the list of common hand gestures. Animations were done for the common hand action units. These were saved into separate files so researchers can use these to animate by connecting each one to another creating words or phrases. This allowed researchers to simply use those animations, and stitch them together to form a gloss. There were signs that were unique which made them must be animated independently.

After connecting the common hand action units, and the unique hand action units to form the glosses, the resulting animations of glosses are used to construct the English sentences.

3.4 Compilation of the Video Elements

Using a video editing software, the different elements of the final output was compiled together into one video. Elements include the animation of the performed signs, the text equivalent in English, and related images.

The text will be synced with the signs using the annotated video as reference. In the text portion of the video, the sentence equivalent will be shown. While the avatar is signing a word, its corresponding exact text translation will be highlighted to allow the viewer to learn the meaning of the sign. Related images are shown as well since it can retain more information thus improve learning (Gutierrez, 2014). Images of the word being translated were drawn by the researchers.

Figure 2: Final Output.

3.5 Evaluation of the output

The evaluation of the final output was done by 5 FSL deaf signers who are 22 years old to 24. They have 5-18 years of experience in signing.

The evaluation sheet was created based on the criteria presented in previous related studies (San-Segundo et al, 2011). The evaluation sheet made was rephrased to simpler English because the evaluators' mode of communication is different from written English. Thus, using simpler English

will be more understandable. The final output was evaluated using a Likert scale in two major categories: animation and performance, wherein each contains subcategories such as the naturalness of the animation, the understandability of the sign, and the usability of the output as a learning material. The evaluation sheet also contains a portion for the evaluator's comments and suggestions.

Rate from 1-6 with 1 (lowest) and 6 (highest).	1	2	3	4	5	6
Animation						
1.The animation is accurate.						
2.The animation is natural looking.						
3.I understand the performed signs.						
4.The avatar is eye-catching and child friendly.						
Performance						
1.Layout of video is organized.						
2.I would use this as a learning reference material.						
Overall Rating						
Comments						

Table 3: Final Output Evaluation Sheet

4. Results and Discussion

The evaluation was done by 5 FSL deaf signers with 5-18 years of experience in signing. The results are as follows.

The first statement, "The animation is accurate.", received an average score of 3.4 out of 6. This may be due to animation errors. There are some signs that the avatar performed incorrectly such as the "Boastful" sign. The sign of "Boastful" is palm rolled with your thumb pointing at yourself. However, the avatar's thumb wasn't in a pointing position. It was mentioned by the evaluators that different FSL schools teaches sign language differently. Another instance was the sign of "Hare". It was mentioned by the evaluator that, in hand sign of "Hare", the fingers were supposed to be bending halfway instead of fully folded. One of the evaluators stated that the signing in the reference video is different from what they are currently signing. It appears that different schools teach sign language differently.

The second statement, "The animation is natural looking.", received a 3.4 out of 6 as well. The low score may be caused by the avatar having no facial expressions. Facial expressions are essential for sign language to express their emotions. If facial expressions and body movements were added, it can possibly increase the score. By adding these, it can make the avatar possess human-like characteristics. An evaluator mentioned that children can't understand if there are no facial expressions involve in the character who is signing.

The third statement, "I understand the performed signs.", received the lowest rating 3 out of 6. This can be due to unclear animations. There were parts of the animation moving too quick or was distorted. One example was from the sign of "Ready". The arm of the avatar during this sign was raised weirdly. Most of the evaluators also commented that the hand movements were unclear and confusing.

The fourth statement, "The avatar is eye-catching and child friendly,", received a relatively low score of 3.2 out of 6. Although the avatar looks pleasant, majority of the evaluators suggested to use characters that looks younger (preferably child) and Filipino to make the children relate to the avatar more. According to one of the FSL deaf signer, since this was meant to be a learning material, it might be better if the signing speed of the avatar is slowed down by 50-60% to ensure that each sign is fully seen.

The fifth statement, "Layout of video is organized.", received an average rating of 3.6 out of 6. There were mixed comments about the layout of the output. Some people suggested to enlarge the avatar to make the hand motions more obvious for the children. There were also comments about slowing down the captions as well for smoother learning and enlarging and adding more color and life to the illustrations. Lastly, there was a comment suggesting to simply just remove the captions because it was very distracting. He added that some deaf children ignore the texts and focus only on the sign language. However, he also added that the vocabulary words can be added either at the start or at the end of the video.

Lastly, the sixth statement, "I would use this as a learning reference material.", received a very high rating of 5.2 out of 6. Although the previous categories received a low rating as compared to the sixth statement, many are fond of the concept of the final output which makes the score high.

Criteria	Average Rating
The animation is accurate.	3.4
The animation is natural looking.	3.4
I understand the performed signs.	**3.0**
The avatar is eye-catching and child friendly.	3.2
Layout of video is organized.	3.6
I would use this video as learning reference.	**5.2**

Table 1: Results of Evaluation

After gathering the data, the evaluators suggested to modify and revise some parts of the animation as the sign language from the video used has a slight difference compared to the sign language they are currently using.

5. Conclusions and Recommendations

Based on the evaluation done, the animation/avatar (1) lacks accuracy, the sign language performed by the person on the video seems to be slightly different than that of the sign language used by the evaluators. This suggest that sign language is very diverse, it does not have a fixed structure; (2) lacks naturalness, it is highly recommended to put facial expressions and body movements when animating. This allows the children to understand more since they are still in the learning age; (3) needs a change of appearance, it is recommended to use or create avatars that looks similar to

the people of the chosen language. This makes the children more entertained, attached and focused on the avatar; (4) needs to slow down by 50-60%, children can't process sign language quick unlike adult signers since they just started learning. Alongside with the animation, it is also recommended to input English words to help boost the vocabulary skills of the children either at the beginning or ending of the video. Overall, the concept of this project is well-received by the evaluators.

6. Acknowledgements

This research would not be possible without the help of FSL deaf signers.

The researchers would like to thank Ms. Archina M. Landayan and Mr. Ismael Somera from De La Salle-College of St. Benilde-School of Deaf Education and Applied Studies for their participation in the annotation of the video.

The researchers would also like to thank Ms. Erika Lauren Aguillon, Mr. Gabino F. Cabanilla II, Mr. Vince Dane R. Lalu, Ms. Ginellyn Mercado, and Ms. Archina M. Landayan for their valuable insights during the evaluation of the final output.

7. Bibliographical References

Ang, K., Antonio, J., Sanchez, D., Yu, S. and Ong, E. (2010). Generating Stories for a Multi-Scene Input Picture. In Proceedings of the 7th National Natural Language Processing Research Symposium, 21-26, November 19-20 2010, De La Salle University, Manila.

Alice. (2017). Retrieved from https://www.alice.org/

Blender Foundation. (n.d.). Blender: About. Retrieved from https://www.blender.org/about/

Brour, M. & Benabbou, A. (2018). ATLASLang MTS 1: Arabic Text Language into Arabic Sign Language Machine Translation System. Retrieved from: https://www.sciencedirect.com/science/article/pii/S1877050919300699?via%3Dihub

Flores, M.F. (2012, July 8). K-12 to use sign language as mother tongue for deaf. Retrieved from Yahoo! Philippines: https://ph.news.yahoo.com/ blogs/the-inbox/k-12-sign-language-mother-tongue-deaf-143729869.html

Gutierrez, K. (2014, July 8). Studies Confirm the Power of Visuals in eLearning. Retrieved from https://www.shiftelearning.com/blog/bid/350326/studies-confirm-the-power-of-visuals-in-elearning

Halawani, S.M. (2008). Arabic sign language system on mobile devices. International Journal of Computer Science and Network Security, 8(1), 251-256.

Huff, E. (2012, June 15). How to help a deaf child to be a successful leader. Retrieved on March 4, 2019 from Reading Horizons: https://athome.readinghorizons.com/blog/ how-to-help-a-deaf-child-become-a-successful-reader.

Imperial, M. F. (2015, June 28). Kinds of sign language in the Philippines. Retrieved on February 13, 2019 from http://verafiles.org/articles/kinds-sign-language-philippines

Jemni, M., Ghoul, O. E., Yahia, N. B., & Boulares, M. (2007). Sign language MMS to make cell phones accessible to deaf and hard-of-hearing community. Conference & Workshop on Assistive Technologies for People with Vision & Hearing Impairments. doi: 10.1145/2207016.2207049

Landayan, Archina (January 24, 2017). The tortoise and the hare in Filipino Sign Language. Retrieved from https://youtu.be/2g6DNncMfms

Lönneker, B., Meister, J. Story Generators: Models and approaches for the generation of literary artefacts. In The 17th Joint International Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing. Conference Abstracts, 126-133, Victoria, BC, Canada, June 2005. Humanities Computing and Media Centre, University of Victoria.

Magee, Paula M., "Challenges with Literacy Development in Children who are Deaf or Hard of Hearing" (2014). Research Papers. Paper 509. Retrieved from http://opensiuc.lib.siu.edu/gs_rp/509

Martinez, L. & Cabalfin, E.P.. (2008, November). Sign Language and Computing in a. Developing Country: A Research Roadmap for the next two decades in the Philippines. Retrieved from https://www.aclweb.org/anthology/Y08-1046

McKay, H., and Dudley, B. (1996). About storytelling: A practical guide. Sydney: Hale & Iremonger.

Mendoza, A. (2018, October 29). The sign language unique to deaf Filipinos. Retrieved on March 4, 2019 from CNN Philippines: http://nine.cnnphilippines.com/life/culture/2018/10/26/Filipino-Sign.Language.html

National Statistics Office. (2003). 2000 census of population and housing. Manila, Philippines: National Statistics Office.

San-Segundo, R., Montero, J.M., Cordoba, R., Sama, V., Fernandez, L.F., D'Haro, L.F.,… Garcia, A. (2011). Design, development and field evaluation of Spanish into sign language translation system. Retrieved from https://core.ac.uk/download/ pdf/12002348.pdf

Shapiro,A. (2015). SmartBody Manual. Retrieved from http://smartbody.ict.usc.edu/HTML/SmartBodyManual.pdf

Tacchetti, M. (2018). User Guide to ELAN Linguistic Annotator. Retrieved from https://www.mpi.nl/corpus/manuals/manual-elan_ug.pdf

Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 45–52
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

LSE_UVIGO: A Multi-source Database for Spanish Sign Language Recognition

Laura Docío-Fernández[1], José Luis Alba-Castro[1], Soledad Torres-Guijarro[1], Eduardo Rodríguez-Banga[1], Manuel Rey-Area[1], Ania Pérez-Pérez[1], Sonia Rico-Alonso[2], Carmen García-Mateo[1]

[1]atlanTTic Research Center for Telecommunication Technologies, [2]Ramón Piñeiro Centre for Research in Humanities
[1]Escola de Enxeñaría de Telecomunicación. Universidade de Vigo, [2]Xunta de Galicia
ldocio@gts.uvigo.es, jalba@uvigo.es, soledadtorres@uvigo.es, erbanga@uvigo.es, mreyarea@gts.uvigo.es, aniaperezperez@uvigo.es, sricalo@cirp.es, carmen.garcia@uvigo.es

Abstract

This paper presents LSE_UVIGO, a multi-source database designed to foster research on Sign Language Recognition. It is being recorded and compiled for Spanish Sign Language (LSE acronym in Spanish) and contains also spoken Galician language, so it is very well fitted to research on these languages, but also quite useful for fundamental research in any other sign language. LSE_UVIGO is composed of two datasets: LSE_Lex40_UVIGO, a multi-sensor and multi-signer dataset acquired from scratch, designed as an incremental dataset, both in complexity of the visual content and in the variety of signers. It contains static and co-articulated sign recordings, fingerspelled and gloss-based isolated words, and sentences. Its acquisition is done in a controlled lab environment in order to obtain good quality videos with sharp video frames and RGB and depth information, making them suitable to try different approaches to automatic recognition. The second subset, LSE_TVGWeather_UVIGO is being populated from the regional television weather forecasts interpreted to LSE, as a faster way to acquire high quality, continuous LSE recordings with a domain-restricted vocabulary and with a correspondence to spoken sentences.

Keywords: Spanish Sign Language (LSE), sign language recognition (SLR), LSE_UVIGO, LSE_Lex40_UVIGO, LSE_TVGWeather_UVIGO, Microsoft Kinect v2

1. Introduction

Automatic speech recognition is one of the core technologies that facilitate human-computer interaction. It can be considered a mature and viable technology and is widely used in numerous applications such as dictation tools, virtual assistants and voice controlled systems. However automatic sign language recognition (SLR) is far less mature.

Some reasons for this have to do with the multimodal nature of sign languages, where not just hands, but also face, head, and torso movements convey crucial information. Others are related with the high number of structural primitives used to build the messages. For example, Spanish spoken language has between 22 and 24 phonemes, but Spanish Sign Language (LSE) has 42 hand configurations, 24 orientations (6 of fingers times 4 of palm), 44 contact places (16 in the head, 12 in the torso, 6 in the dominated hand/arm and 10 in space), 4 directional movements and 10 forms of movement (according to (Herrero Blanco, 2009), although there is no unanimity in this classification, see for example CNSE (2008)).

The study of the state of the art suggests that machine learning applied to SLR will be sooner or later able to overcome these difficulties as long as there are adequate sign language databases. Adequate means, in this context, acquired with good quality, carefully annotated, and populated with sufficient variability of signers and visual contexts to ensure that the recognition task is robust to changes in these factors.

Unfortunately, only a few sign languages offer linguistic databases with sufficient material to allow the training of complex recognizers (Tilves-Santiago et al., 2018; Ebling et al., 2018), and LSE is not one of them. There have been some efforts to collect the variety of LSE signs through different recording technologies and with different purposes. The video dataset from Gutierrez-Sigut et al. (2016) contains 2400 signs and 2700 no-signs, grammatically annotated, from the most recent standardized LSE dictionary (CNSE, 2008). Even though this controlled dataset is very useful to study the variability of Spanish signs, the poor variability of signers (a man and a woman signing half dictionary each), the absence of inter-sign co-articulation and the small resolution of the body image, precludes it from its use for training machine learning models for signer-independent continuous Spanish SLR.

The Centre for Linguistic Normalization of the Spanish Sign Language (CNLSE, acronym in Spanish) has been developing a corpus for years in collaboration with numerous associations and research centres in the state. It is composed of recordings of spontaneous discourse, very useful to collect the geographical, generational, gender and type of sign variation of the LSE. However it is not appropriate for SLR training in a first phase, which would require a database with a high number of repetitions per sign and, probably, the temporal segmentation of the signs collected in the recordings.

A completely different LSE dataset (Martinez-Hinarejos, 2017) was acquired with the Leap Motion infrared sensor that captures, at a short distance, the position of the hands and fingers, similarly to a data glove but touchless. This publicly available dataset is composed of a main corpus of 91 signs repeated 40 times by 4 people (3640 acquisitions) and a 274 sentences sub-corpus formed from 68 words of

the main corpus. The technology of Leap Motion limits its use to constrained movements (close to the device and without self-occlusions) and prevents capturing arms, body motion and facial expressions. Therefore, its usefulness to SLR would probably be limited to fingerspelling.

From this review we conclude the need to build a new multi-source database, which we will call LSE_UVIGO, specifically designed to support our ongoing research on SLR, and that of others. Our team is made up of two research groups of the University of Vigo: the Multimedia Technology Group (GTM) and the Grammar, Discourse and Society group (GRADES). GTM has accredited expertise on facial and gesture analysis, and speech and speaker recognition, and GRADES has a longstanding expertise on LSE and interaction with deaf people. With the development of LSE_UVIGO we intend to support fundamental and applied research on LSE and sign languages in general. In particular, the purpose of the database is supporting the following or related lines:

- Analyse the influence of the quality of video footage on the processing of the video stream for segmentation, tracking and recognition of signs.
- Quantify the advantages of including depth information.
- Segment and track upper-body parts in 2D/3D, and quantify the benefits of an accurate segmentation on SLR.
- Develop tools to align continuous speech and LSE.
- Develop signer-independent sign to text/speech translation, both word-based and sentence-based, including fingerspelling.
- Analyse the influence of face expression and body movements on decoding sign language sentences.
- Measure the robustness of sign language processing modules against changes in the scenery.

2. LSE_UVIGO Database

Initially, LSE_UVIGO consist of two different datasets that complement each other to the above purposes: the LSE_Lex40_UVIGO and the LSE_TVGWeather_UVIGO. The first one is intended to support research on LSE through high quality RGB+D video sequences with high shutter speed shooting. The second one is composed of broadcast footage of the weather forecast section in Galician Television (TVG) news programs. Following sections explain with more detail both datasets.

2.1 LSE_Lex40_UVIGO Dataset

This subset is a multi-sensor and multi-signer dataset acquired from scratch. It is thought as an incremental dataset, both in complexity of the visual content and in the variety of signers, most of them deaf.

LSE_Lex40_UVIGO is intended to cover most of the necessities of the research community working in SLR: static and co-articulated sign recordings, both fingerspelled and gloss-based isolated words, and sentences. The recording is done in a controlled lab environment in order to obtain good quality videos with sharp video frames and RGB and depth information, making them suitable to try different approaches to automatic recognition. The RGB and depth information are co-registered in time which allows researchers to work not only on recognition, but also on tracking and segmentation.

In its present form, the contents of LSE_Lex40_UVIGO are organised in three sections:

- The LSE alphabet, composed of 30 fingerspelled letters.
- 40 isolated signs, which can be static or dynamic, in which one or both hands intervene, with symmetric or asymmetric movement, and with different configurations, orientations and spatial-temporal location. They were selected according to linguistic-morphological criteria so as to reflect different modes of articulation that may affect the complexity of SLR (Torres-Guijarro, 2020).
- 40 short sentences related to courtesy and interaction. The sentences were chosen based on vocabulary that is traditionally included in introductory LSE courses. Each sentence ranges from one to five signs in length.

In order to facilitate the labelling process, the signs are performed in a standardized way, trying to avoid dialect variations of glosses as much as possible.

Recording Software and Setup

The UVigoLSE_Lex40 dataset is being recorded with two visual devices: a Microsoft Kinect v2, which captures both RGB video (resolution 1920x1080 pixels @30 FPS) and depth maps (resolution 512x424 pixels @ 30 FPS), and a Nikon D3400 camera which captures high quality RGB video signals (resolution 1920x1080 @ 50 FPS). The shutter speed of the Nikon camera is set to 1/240 sec. to freeze the movement of the signing sequence even for quite fast movements of the signer. Both devices are fitted on a rigid mount on a steady tripod. The mount is placed in front of the signer facing the signing space, and the recording location has been carefully designed to facilitate the framing, focusing, lighting and setting the distance to the signer. Figure 1 shows the recording setting.

Figure 1: Set up of the dataset acquisition. Kinect and Nikon devices are rigidly mounted on a tripod at a fixed distance to the signer, that is uniformly illuminated over a somehow uniform background (location settings vary). No restrictions on clothing are imposed.

To facilitate the introduction of the metadata of the recording session (date and place, operator, recording devices) and the signer self-reported information both written and signed (name, sex, year of birth, school, dominant hand, place of residence, hearing/deaf, at what age she/he started learning LSE, and at what age she/he went deaf), an acquisition platform has been programmed in MatLab®, which also allows simultaneously recording from the two devices.

2.2 LSE_TVGWeather_UVIGO Dataset

Nowadays it is nearly impossible to acquire a large-scale, high-quality LSE dataset which captures all the difficulties of the SLR task. The main reason for this is the high cost of designing, recording and annotating a dataset with a large vocabulary and a sufficient number of signers. To solve this issue, public video sources available in LSE can be used, such as websites dedicated to teaching sign language, and TV programs interpreted in LSE.

Monday through Friday, the midday newscast of the regional television network (TVG) is interpreted in LSE. Both the original broadcast in Galician language and the LSE version, are available on the TVG website. The news domain is too ample for considering the acquisition of a database for continuous SLR. Therefore, inspired by other authors' work (Koller et al., 2015), we decided to focus on a restricted domain: weather forecasts.

LSE_TVGWeather_UVIGO dataset is being populated with weather forecasts from the TVG news on workdays, with a typical duration of 1-2 minutes. The main characteristics of the video codec are: H.264, resolution 1280x720, 50 frames per second. As illustrated in Figure 2, the sign language interpreter occupies about 20% of the image (around 400*470 pixels), a screen portion substantially larger than that used in other TV channels. Every video is automatically annotated at the word level by means of our Galician automatic speech recognizer (ASR) system. This transcription is then manually reviewed at a higher "segment" level (quite similar to a breath-group level) using ELAN, leaving the weather forecast ready for further annotation (as illustrated in Fig. 6; detailed information about annotation is given in Section 4).

3. Video and Depth Signal Post-processing

As explained in Section 2.1, LSE_Lex40_UVIGO recordings are acquired simultaneously with a Nikon camera and a Kinect. The Nikon provides high quality RGB, and the Kinect provides complementary depth information, quite useful for segmenting regions of interest in RGB images, such as hands, arms and face. In the following sections, details are given on the time-alignment of depth and video signals, and on the segmentation process itself. Segmentation will also be applied to LSE_TVGWeather_UVIGO.

Figure 2: weather forecast in the regional TV network (TVG), interpreted to LSE.

3.1 Time-alignment and Transferring of Depth to the RGB Streams

In order to complement Nikon's images with depth information from Kinect, a two-step co-registering and alignment process is needed. This process is outlined in Figure 3.

The first step entails co-registering color and depth from the Kinect sensors. Although RGB and depth information are gathered by the Kinect simultaneously, these two signals are not synchronous because their sensors are initially triggered at different moments, the periodic acquisition has some jitter, and a frame from any of the sensors is occasionally lost. In order to perform a temporal alignment over the whole sequences we have used the skeleton landmarks provided by Kinect software development kit. After calculating the optimal projective transformation between pairs of temporally-aligned frames, we apply a Dynamic Time Warping (DTW) algorithm using the minimum squared error (MSE) of the location of skeleton landmarks among the co-registered pairs as the distance measurement. This last step is avoided if absolute timestamps are preserved during the recording of RGB and depth information[1].

The second step consists in co-registering Kinect RGB and Nikon RGB. In this case we cannot use the Kinect's skeleton landmarks, so we have resorted to OpenPose software to co-locate a set of landmarks in temporally similar frames and calculate a geometrical transformation to co-register the short focal length Kinect RGB+D maps onto the larger focal length Nikon RGB image. Given that the triggering (start, stop and period) and acquisition period are also different, we need to temporarily align the sequences using again a DTW algorithm. Similarly to the previous step, the distance measure between frames is the MSE of the location of OpenPose landmarks.

[1] The first recordings of isolated signs in LSE_Lex40_UVIGO were acquired without the absolute timestamp.

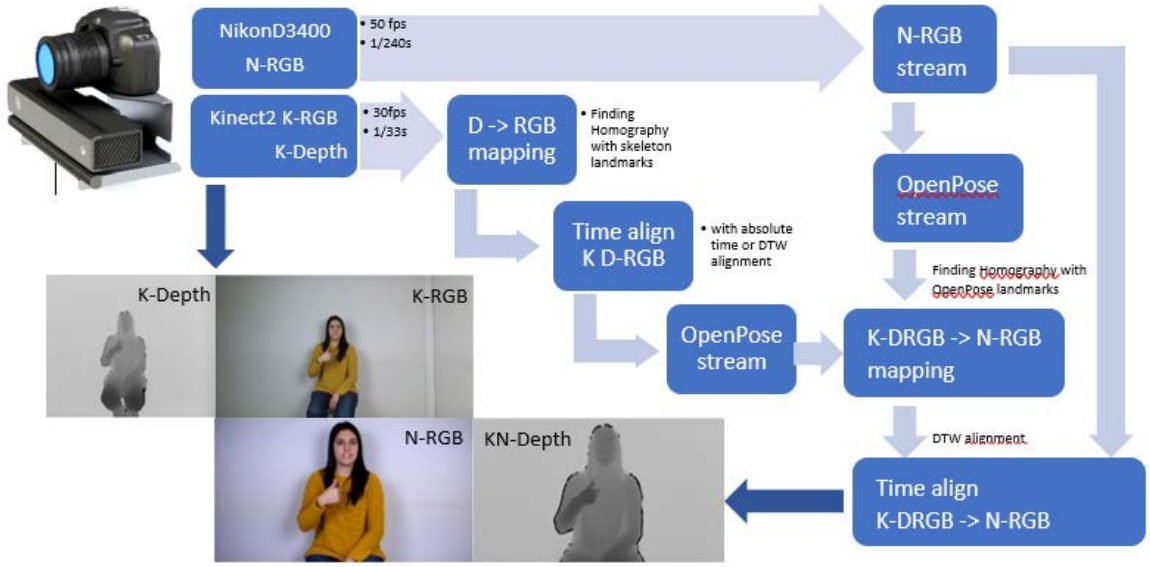


Figure 3: Flow diagram of the post-processing to align all the streams and transfer Kinect depth information to the Nikon acquired RGB stream.

3.1 Hands and Face Segmentation

A recurrent issue in object and instance recognition is the amount of context needed to identify the object or its specific configuration. SLR does not get rid of this problem, and despite some efforts on determining whether perfectly segmented hands and face work better in SLR than the complete image containing the full body context (Huang, 2015; Camgoz, 2017; Koller, 2019), more studies are needed in this field.

Most sign language interpreters use dark clothes to facilitate the contrast of hands over the body, so it seems that an automatic recognition system could benefit from a proper segmentation of the hands. But the relative location of the hands and arms with respect to the body and face is also crucial, so keeping the visual context could help the system. Current techniques using deep neural networks fed by holistic visual appearance seem to digest unsegmented objects properly, but only a large variety of examples (Li, 2019) will help the network to simulate the visual attention made by the brain, and thus to get rid of the non-discriminative surrounding information. Unfortunately, Spanish sign language datasets are still too small to benefit from this approach.

To support research on the influence of segmentation, LSE_UVIGO will also provide a segmentation map, so researchers can directly try their algorithms with or without context information. Figure 4 shows a simplified flow diagram of the image processing, which makes use of colour, OpenPose landmarks of the RGB stream (Cao et al., 2018; Simon et al., 2017), and depth when available. A similar segmentation approach but using just the Kinect sensors was proposed in (Tang, 2014). Image at left shows the result of using a generic skin map. It is clear that colour information alone was not able to eliminate the sweater and the neck information. Picture at right shows the original image filtered by a probability map that takes into account a user-specific skin-map, the depth co-registered image and the distance to the OpenPose landmarks at hands and face. So, instead of providing a final binary mask, we store in the database a probability map with real values between 0 and 1, so researches can choose to threshold at different levels to include more or less body information, or even just use the map as a filter that preserves the information of hands and face and attenuates the rest in a 'saliency-map' way. It is important to highlight that the Kinect RGB stream, as most of the SL videos in other datasets, contains blurred hands when movement is relatively fast because of the shutter speed of 1/33 secs. For this reason we have resorted to the Nikon's streams with shutter speed of 1/240 secs, which allows to freeze most of the very fast hand movements and allows a more accurate segmentation.

4. Database Annotation

We are enriching the database with detailed manual and semi-automatic annotations using the ELAN software package (Brugman & Russell, 2004). The annotation is divided into several parts, similarly to the CORILSE corpus annotation (Cabeza-Pereiro et al., 2016):

4.1 Annotation of Manual Components (MC)

The annotation of MC includes the start and end points of every sign, transition movements between signs and discourse pauses, and the gloss ID with respect to an

annexed lexical database. This annotation phase involves the tiers *MD_Glosa* (Gloss for right hand) and *MI_Glosa* (Gloss for left hand). It is important to highlight that some non-lexical units are also annotated in this phase, the most important one being the buoy (B) hand indicating that one or both hands are paused in a specific position and configuration after (or even before) its participation in a

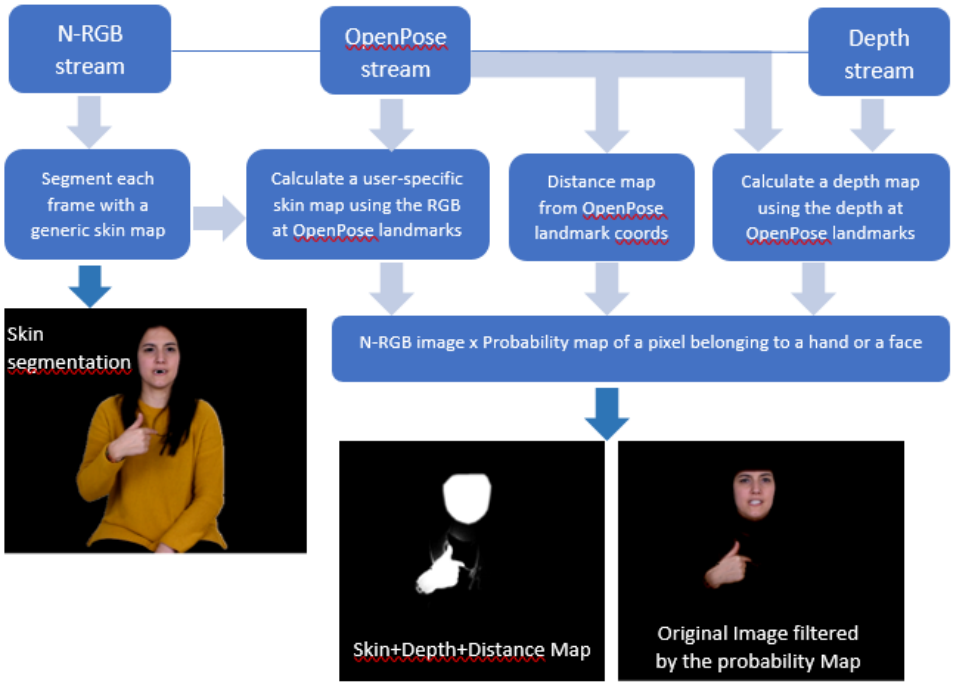


Figure 4: Simplified flow diagram to segment hands and face from the video sequences.

sign. Other non-lexical or semi-lexical units are also annotated like gestures (G) and indexes (INDX) respectively.

4.2 Annotation of Non-Manual Components (NMC)

The annotation of NMC is still under development. The number of components defined by Cabeza-Pereiro et al. (2016) is much larger than needed for the purpose of this database. We will annotate the NMC useful for disambiguation of a sign (like the eyebrows in SWEET and PAIN), those that modulate the discourse (like movement of eyebrows and mouth in a question clause) and those that are modifiers of the sign (like shape of mouth and cheeks when indicating a big amount of people, work, money, etc.). Another type of NMC to annotate is blinking, that helps to determine the end of a clause in LSE. Action Units provided by OpenPose are being used for detecting the NMC in the video stream and will be imported as NMC tiers. Manual revision from an expert LSE signer will be needed to eliminate false positives and add false negatives in these tiers.

4.3 Annotation of Other Linguistic Information

The literal translation to Spanish (tier *Trad*) is annotated, and also a segmentation of each predicative expression (a 'clause-like unit' or CLU, to borrow a term used by Johnston (2013) and Hodge (2013)). Each CLU will have a different reference in the tier *Ref* (Reference) and will facilitate the construction of LSE/Spanish pairs for training and testing end to end translation systems. If the dataset only contains text and signs, as in LSE_Lex40_UVIGO, the *Ref* is aligned with the set of signs that form the CLU (see Fig.5). If the dataset contains also a speech stream simultaneously translated to LSE, as in LSE_TVGWeather_UVIGO, there are two *Ref* tiers; *Ref_LO* for speech CLUs and *Ref_SL* for LSE CLUs. Given that the LSE signer translates from a speech stream in real-time (Galician language in this case), there's a variable amount of time shift between them, so detailed annotation of the spoken-signed CLU pairs is a great help for developing translation systems. Two more tiers are annotated in the LSE_TVGWeather_UVIGO: *Word* and *Segment*. The first one corresponds to the automatic speech recognizer (ASR) output, with timestamps between words, while the second one corresponds to the manual review of the sentences extracted automatically from the sequence of words from the *Word* tier.

Figure 5 shows a screenshot of the annotation of LSE_Lex40_UVIGO dataset and Figure 6 shows the annotation of the LSE_TVGWeather_UVIGO.

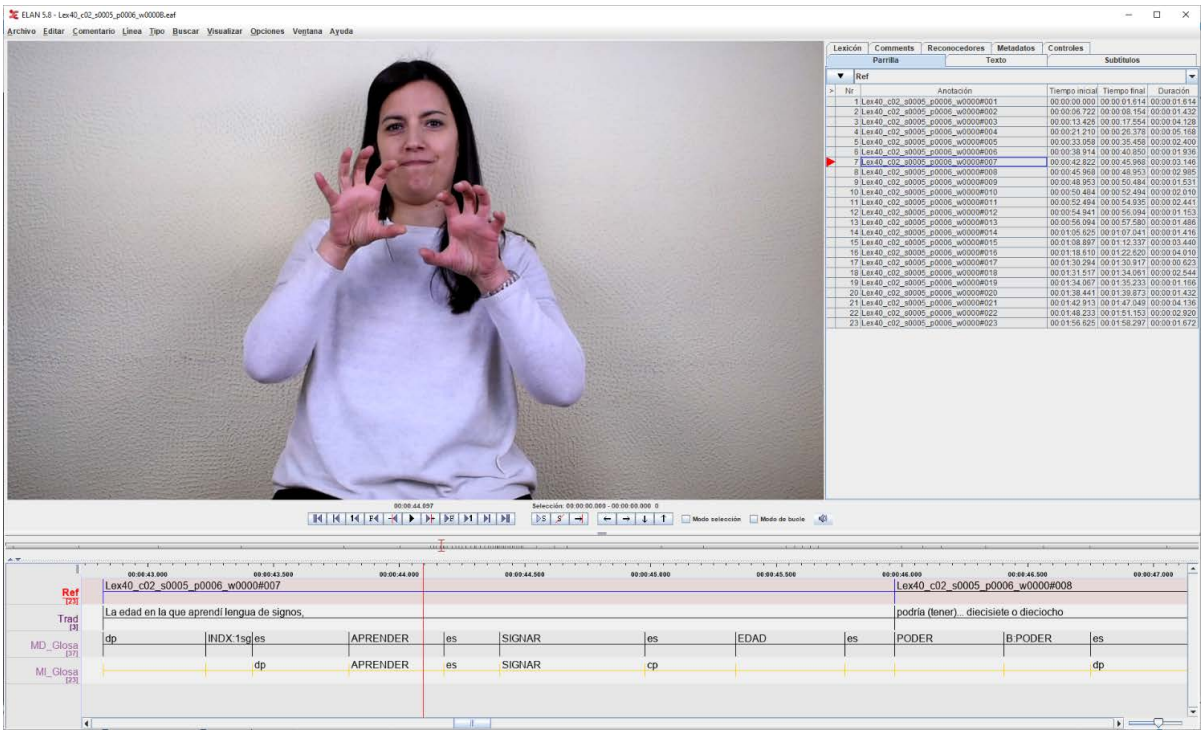


Figure 5: Example of ELAN annotation tiers in LSE_Lex40_UVIGO dataset. *Ref* tier (encapsulates predicative expressions), *Trad* tier (the Spanish translation of the signed sentence), *MD_Glosa* and *MI_Glosa* (the right and left hands lexical signs "APRENDER, SIGNAR, EDAD, PODER"; transitions "dp" -from pause-, "es" -inter sign transition-, "cp" -to pause-; and semi-lexical signs "INDX:1sg" -pointing to subject-, "B:PODER" -buoy sign-).

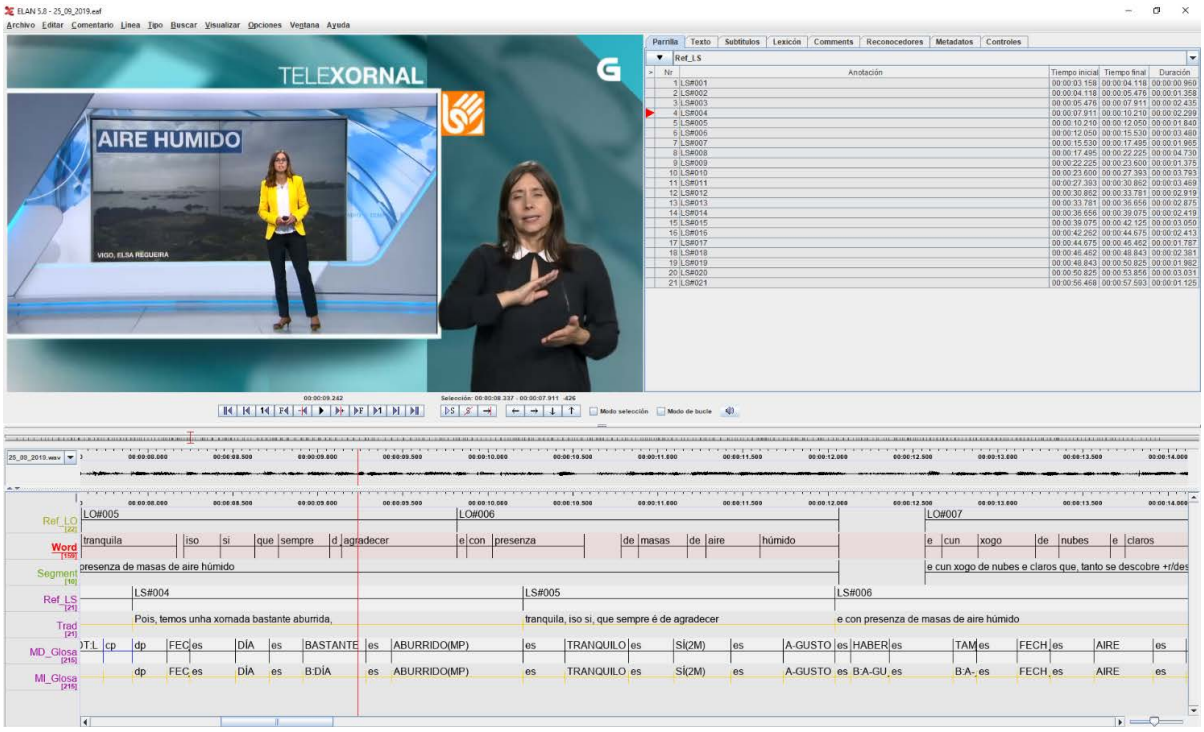


Figure 6: Example of ELAN annotation tiers in LSE_TVGWeather_UVIGO dataset. *Ref_LO* and *Ref_LS* tiers form pairs of spoken-signed CLUs, *Word* and *Segment* tiers come from the ASR and the manual review, respectively, *Trad* tier is the Galician utterance (hopefully quite close to the ASR in the *Word* and *Segment* tiers), but aligned with the LSE stream, and the rest of tiers as in Fig. 5.

5. Current state of the Database and Further Work

We started recording LSE_Lex40_UVIGO in May 2019 and, to this date, 35 signers have contributed to it. They come mostly from the deaf community and display a range of ages and fluency in sign language, and gender parity. So far, most of the videos have been recorded in the Association of Deaf People of Vigo (ASORVIGO), and the rest in the School of Telecommunications Engineering and in the Faculty of Philology and Translation of the University of Vigo. In all three cases the distance to the cameras and the framing was similar, while the background of the image has variations: It is a bare wall painted light in two of the locations, and is covered by a green fabric to eliminate reflections in the third. We did not impose any requirements on signer clothing. In future recordings we will incorporate other locations, lighting conditions and background types to test the robustness of the ASLR against this type of variation in the recording conditions.

Table 2 summarizes the main figures of LSE_Lex40_UVIGO dataset up to now: columns 2 through 5 indicate the number of different items in each section of the dataset (alphabet, isolated signs and sentences), the number of signers that have contributed to each section, the number of available recording of each item, and the total duration of the recordings. We plan to incorporate a new section to the dataset, namely 40 fingerspelled words.

Regarding LSE_TVGWeather_UVIGO dataset, recording started in August 2019 at a rate of about 18-20 videos per month. To this moment, about 100 videos have been recorded, most of which last between 1 and 2 minutes. Usually they are signed by the same person.

We are managing the transfer of the rights of the images by the signers in accordance with the European regulation of the protection of personal data, so a first release of the LSE_UVIGO database may be made available to the research community in the coming months.

Database section	# Items	# Signers	# Recordings	Total recordings duration (hh:mm:ss)
Alphabet	30	3	90	00:03:45
Isolated signs	40	32	1368	01:23:50
Sentences	40	13	493	00:34:46
Total	110	35	1951	02:01:31

Table 2: current state of LSE_Lex40_UVIGO dataset

6. Acknowledgements

This research is funded by the Spanish Ministry of Science, Innovation and Universities, through the project RTI2018-101372-B-I00 *Audiovisual analysis of verbal and nonverbal communication channels (Speech & Signs)*; by the Xunta de Galicia and the European Regional Development Fund through the Consolidated Strategic Group atlanTTic (2016-2019); and by the Xunta de Galicia through the Potential Growth Group 2018/60.

The authors wish express their immense gratitude to the Association of Deaf People of Vigo (ASORVIGO) and the Federation of Associations of Deaf People of Galicia (FAXPG) for their collaboration in the recording of the database LSE_Lex40_UVIGO.

7. References

Brugman, H. & Russell, A. (2004). Annotating multimedia/multi-modal resources with ELAN. Paper presented at the LREC 2004, In Proceedings of the Fourth International Conference on Language Resources and Evaluation, Lisbon, Portugal.

Cabeza-Pereiro, M. C., Garcia-Miguel, J. M., García-Mateo, C., & Alba-Castro, J. L. (2016). CORILSE: a Spanish Sign Language Repository for Linguistic Analysis. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (pp. 1402-1407).

Camgoz, N. C., Hadfield, S., Koller, O., & Bowden, R. (2017). Subunets: End-to-end hand shape and continuous sign language recognition. In 2017 IEEE International Conference on Computer Vision (ICCV), pp. 3075-3084.

Cao, Z., Hidalgo, G., Simon, T., Wei, S. E., & Sheikh, Y. (2018). OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. arXiv preprint arXiv:1812.08008.

CNSE Foundation (2008). Diccionario normativo de lengua de signos española: Tesoro de la LSE [DVD].

Ebling, S., Camgöz, N. C., Braem, P. B., Tissi, K., Sidler-Miserez, S., Stoll, S., ... & Razavi, M. (2018). SMILE Swiss German sign language dataset. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

Gutierrez-Sigut, E., Costello, B., Baus, C., & Carreiras, M. (2016). LSE-sign: A lexical database for spanish sign language. Behavior Research Methods, 48(1), 123-137.

Herrero Blanco, Á. L. (2009). Gramática didáctica de lengua de signos española, LSE. Ediciones SM, Madrid.

Hodge, G. (2013). Patterns from a signed language corpus: Clause-like units in Auslan (Australian sign language). Ph.D. thesis, Sydney: Macquarie University.

Huang, J., Zhou, W., Li, H., & Li, W. (2015, June). Sign language recognition using 3d convolutional neural networks. In 2015 IEEE international conference on multimedia and expo (ICME) (pp. 1-6). IEEE.

Johnston, T. (2013). Auslan Corpus Annotation Guidelines. Retrieved from http://media.auslan.org.au/attachments/Johnston_AuslanCorpusAnnotationGuidelines_February2016.pdf

Koller, O., Forster, J., & Ney, H. (2015). Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers.

Computer Vision and Image Understanding, 141, 108-125.

Koller, O., Camgoz, C., Ney, H., & Bowden, R. (2019). Weakly Supervised Learning with Multi-Stream CNN-LSTM-HMMs to Discover Sequential Parallelism in Sign Language Videos. IEEE transactions on pattern analysis and machine intelligence. doi: 10.1109/TPAMI.2019.2911077

Li, D., Rodriguez-Opazo, C., Yu, X. and Li, H. (2019). Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison. Accepted at IEEE 2020 Winter Conference on Applications of Computer Vision (WACV '20), March 2020. https://arxiv.org/abs/1910.11006

Martínez-Hinarejos, C. D., & Parcheta, Z. (2017). Spanish Sign Language Recognition with Different Topology Hidden Markov Models. In INTERSPEECH (pp. 3349-3353).

Simon, T., Joo, H., Matthews, I., & Sheikh, Y. (2017). Hand keypoint detection in single images using multiview bootstrapping. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (pp. 1145-1153).

Tang, A., Lu, K., Wang, Y., Huang, J., & Li, H. (2015). A real-time hand posture recognition system using deep neural networks. ACM Transactions on Intelligent Systems and Technology (TIST), 6(2), 1-23.

Tilves-Santiago, D., Benderitter, I., & García-Mateo, C. (2018). Experimental Framework Design for Sign Language Automatic Recognition. In IberSPEECH (pp. 72-76).

Torres-Guijarro, S., García-Mateo, C., Cabeza-Pereiro, C., Docío-Fernández, L. (2020). LSE_Lex40_UVIGO Una base de datos específicamente diseñada para el desarrollo de tecnología de reconocimiento automático de LSE. Revista de Estudios de Lenguas de Signos (REVLES), 2.

Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 53–60
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

Elicitation and Corpus of Spontaneous Sign Language Discourse Representation Diagrams

Michael Filhol
Université Paris–Saclay, CNRS, LIMSI
Campus universitaire, bât. 507
Rue du Belvédère
91400 Orsay
michael.filhol@limsi.fr

Abstract

While Sign Languages have no standard written form, many signers do capture their language in some form of spontaneous graphical form. We list a few use cases (discourse preparation, deverbalising for translation, etc.) and give examples of diagrams. After hypothesising that they contain regular patterns of significant value, we propose to build a corpus of such productions. The main contribution of this paper is the specification of the elicitation protocol, explaining the variables that are likely to affect the diagrams collected. We conclude with a report on the current state of a collection following this protocol, and a few observations on the collected contents. A first prospect is the standardisation of a scheme to represent SL discourse in a way that would make them sharable. A subsequent longer-term prospect is for this scheme to be owned by users and with time be shaped into a script for their language.

Keywords: Sign Language, Spontaneous representation, Writing system

1. Introduction

Sign Languages (SLs) are gestural oral languages, with no written form (Pizzuto and Pietrandrea, 2001). A few systems have been proposed to equip Sign with a standard script, some for a particular SL, others claimed for any SL (Grushkin, 2017; Kato, 2008; McCarty, 2004; Lessa-de Oliveira, 2012; Barros, 2008). SignWriting (Sutton, 2014) is the most well known, and has been at the heart of a few research projects, e.g. to test its use in educational environments, or to equip it with a software editor. HamNoSys (Prillwitz et al., 1989; Hanke, 2004) is the most notoriously known in the computer science domain because it has successfully been processed as input to synthesise signs with a virtual signer (3d avatar) (Elliott et al., 2004; Elliott et al., 2008).

However, none has come at all close to being shared and practiced by large numbers of users with ease, nor does any seem to be currently gaining momentum. Whatever the status of those systems at this point, it remains that Sign Languages have no accepted writing system. But this does not mean that it is impossible to note, draw, graph, or sketch out anything that represents discourse in SL. Indeed, SL users come up with solutions when they need a functional equivalent to writing. The goal of this paper is to study them.

Of course, many simply write text in an official or locally dominant language they happen to know. But this means using the writing system of a different, "foreign" language, in other words translating into a separate one entirely, which falls out of our scope. Instead of proper verbal sentences, articles may otherwise be dropped, arrows used between two written clauses, or pictograms drawn instead of spelling out words. Comparable to shorthand note taking, these techniques do part from the full set of syntactic constraints. However, they remain tied to the written language, constrained by its vocabulary and inspired by its canonical linear order, e.g. subject-verb-object.

In the past years, we encountered a variety of different hand-scripted productions aimed at capturing SL on paper in order to be read later, without or with negligible support from a different written language. Each language user producing their own personal approach that was neither learnt nor theorised as a system, we call those "spontaneous" representations of the SL discourse represented in their contents.

In the next section (§2.), we present a variety of cases and examples of such productions. After explaining their potential scientific value, section 3. defines an elicitation protocol to collect a corpus of similar data, enabling statistical analyses. Section 4. reports on the current state and amount of collected data, and lists a few observations we were able to draw from it.

2. Spontaneous Sign Language representation

We found three different types of situations where SL was spontaneaously captured in representations on paper by language users.

First, many SL users **preparing signed speeches** have been found to use graphic support to represent the discourse to be delivered. What is more, long discourse has more than once been seen drawn on a paper feed, used as a teleprompter at the time of signing, e.g. in front of a camera. This was done either by scrolling the paper down as the speech was delivered, or by playing a video of the paper feed, itself filmed beforehand in a slow camera travel down from the top. To these users, graphical schemes could always be found sufficient to express the whole of the production, and are naturally preferable to text because they are in direct relationship with the signing space, though with one fewer spatial dimension. An example of hand-drawn teleprompter scroll is given in fig. 1.

The second case of spontaneous diagrams we wish to report on comes from an interesting position at the *Institut*

Figure 1: Diagram feed for a "teleprompter" scroll

National des Jeunes Sourds[1] (INJS) in Paris, a historic deaf school where teachers teach classes in LSF[2] today. At INJS, we met teachers upholding that SL natives should be able to write and turn in homework in a written form in their language, i.e. the one they think and organise ideas in. This implies not to film themselves but produce handwritten work on paper, and not to require written French but make SL the represented language. Since there is no full writing system for SL, these teachers ask their students to draw SL the way they feel it should, provided they can understand the signing that motivated the drawing as they read it. The school has kindly agreed to share a few of those productions with us. Figure 2 shows one of the pages of a piece of homework.

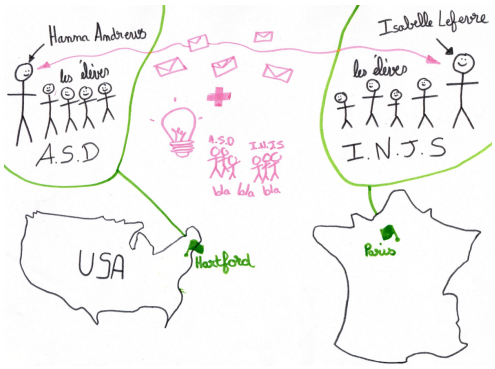


Figure 2: Diagram produced by an INJS pupil

The third use case can be observed in the domain of text-to-Sign translation. Professionals draw "deverbalising diagrams" of the source text in a first stage of their process, to represent its full meaning in a graphical form (Seleskovitch and Lederer, 1985; Athané, 2015). It enables them to work further from the diagrams alone, leaving the source texts aside and avoiding the translation bias they could induce (e.g. sentence order or lexical choices). Figure 3 is an example of a deverbalising diagram, representing a source text of 99 words.

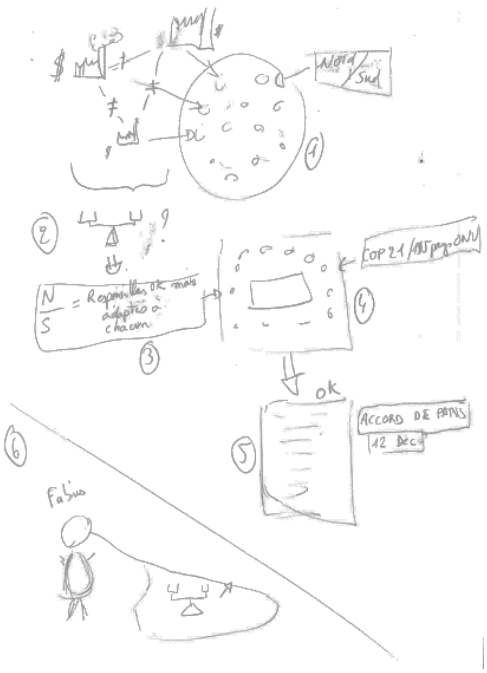


Figure 3: Deverbalising diagram produced in a text-to-SL

After a few such diagrams shared with us, some drawn symbols or higher-level graphical layouts appeared strikingly consistent:

[1] National institute for deaf youth.

[2] *Langue des signes française* = French Sign Language.

- in meaning, e.g. specific style of arrows for displacements;
- or in the forms produced to read them, e.g. relative positions in signing space similar to those in the diagrams.

Moreover, these consistent choices could be seen not only in the diagrams of every person, but also in the whole set of diagrams, across authors. This indicated that even without a designed system, SL users produce sponteneously similar diagrams to some extent.
Unfortunately, these are usually productions discarded after use, considered as private intermediate steps not worth keeping. Because it is the only intended result, authors would rather keep a video recording of the final speech produced than personal drawings which only they at best would make sense of afterwards. But to us, the regularities found in those spontaneous productions constitute a window on the way authors internally work the cogs of their own language, more or less consciously separating them out and linking them together in diagrams.
We hypothesise that patterns can be found in these diagrams, and formalised, and that a lot can be learnt from them. To test this, we decided to build a corpus of such productions to enable their analysis. The next section justifies and explains the protocol we designed to elicit and collect diagrams of the kind spontaneously drawn by SL users representing SL discourse.

3. Corpus elicitation

3.1. Elicitation protocol

Before asking participants to produce data, a formal specification of the tasks and selection of the elicitation material was required to maximise useful contents. However, the point was to elicit spontaneous drawings in order to observe the regularities emerging from language users with as little bias as possible in the productions. So we had to avoid making the informants self-conscious about any lack or consistency of an underlying system in their own diagrams. In other words, in the design of the elicitation task and material, we wanted to control variables observed or likely to impact the drawings, but not to affect how the informants approach composing the drawings themselves.
This to us implied asking the participating subjects to **draw for themselves**, and not for an unknown third party to read. The reason is that since there is no standard, existing system for the diagrams, nothing too specific can be encoded in a diagram and assumed to be fully understood from the diagram alone. Drawing for anybody else at this time forces to use more generic conventions and icons not to turn the productions into a form of guessing game. But on the contrary, we wish to collect as many examples as we can of these personal, possibly creative graphics, letting the authors mirror their (more or less conscious) idea of the internal workings of their language.
Secondly though, if we allow for personal encodings of specific constructs, meanings or gestures, there is a chance we might not understand what they stand for. So if we eventually want to be able to interpret everything of the diagrams and compare them to how they were intended to be read, we must also ask the authors to **deliver the signed version** of what they captured in their drawn productions. With this, any part of the diagram should be interpretable with the help of the signed counterpart.
A third concern then is that if one draws to produce the signed version immediately afterwards, one can easily use short term memory and reduce the information load in the diagram, expanding it as they sign the final discourse. In a written/vocal language, this is analogous to a debater jotting down talking points before developing a rebuttal. Short-term memory stores the contents of the intended speech instead of the writing, which only serves to cue the reader. To enforce the full content of the discourse in the diagram, participants should be drawing with the thought that they will have to **read them later**. Trading off between draw–read separation time and participant time commitment, we told them that they would have to produce a signed read-out of their diagrams two weeks after drawing them.
This in turn raises the new concern of participants not being able to read from their diagrams or, if proactive about it, overloading the diagrams with excessive detail or written language. We compensated this with two further decisions for the protocol. The first one is to allow participants to work on the signed productions immediately they turn their diagrams in (but without telling them in the first place). This way, anything included in the drawings that would have been unclear later can be delivered relying on short-term memory. As long as they do not expect it while they are composing the diagrams, this is acceptable since we are not testing the SL production but only using it to help in later diagram interpretation and analyses. The second decision is to tell them from the start that they will have to read from their diagrams only, but that they will be allowed to fall back on source elicitation documents if they are missing elements to complete the discourse, provided they report those cases.
To summarise, we tell the participants in the beginning that: (1) they are drawing for themselves; (2) they will deliver diagrams first, and a signed version reading off of them two weeks later; (3) should they not understand their diagram, they will be allowed to go back to the source elicitaion material. When they deliver the diagrams, they are told that they may work on the signed version straight away.
Lastly, we decided to allow people to download the elicitation material and work from home, organising the time spent on drawing as they pleased. This was to facilitate the experiment in terms of logistics and comfort for the participants. Besides, a controlled environment would not necessarily add more reliability to the data than it would distance the productions from what they normally shape into. For example, people might need to look up information online, or to take breaks between tasks for better attention and results (no expediting drawings only to finish in an artifical time frame).

3.2. Controlled variables

Many parameters are likely to impact such drawings, which makes it difficult to control the variables entirely. However,

we have listed a few major ones, likely or observed to create different styles, approches or choices in the collected diagrams, and for which a control is possible. We present them in the subsections below.

Form/meaning source

A language is a communication system associating forms (observed or produced) to meaning (interpreted or intended, repectively). A writing system usually captures elements of either form or meaning for a most part, although always of both, in imbalanced proportions. For example in the Spanish writing system, the letter "M" captures the production of the (phonological) form /m/, while the "¿ ... ?" character pair captures both the meaning of interrogation and the form of changing the pitch when pronouncing the enclosed contents. The scripting strategy capturing form is called phonography; the one capturing meaningful units is called logography.

In a spontaneous diagram task, i.e. without an existing system constraining the drawn production, it is likely that the productions will depend on what the user is exposed to or has knowledge of before drawing. For example, a person taking notes of a delivered lecture is exposed to already formed articulations, before even interpreting its meaning, though both can be considered out of her control. On the opposite end, a speaker authoring content from scratch is exposed to neither before drawing. The case of translation presented above is interesting to consider here, as its essence is to put together a non-given target form, while preserving a given source meaning. It is therefore an intermediate case. These three cases are summarised and labelled in table 1, in decreasing order of form/meaning exposure before drawing.

	form given	meaning given
stenographer	yes	yes
translator	no	yes
author	no	no

Table 1: Form/meaning exposure profiles

Without a standard drawing system, these differences in exposure before drawing will potentially inflict on the drawn productions, in ways and to measures still unknown. We therefore wanted to elicit diagrams in each of those situations. One of the first questions after the collection would be to situate the spontaneous diagrams with respect to the phonography–logography distinction, perhaps depending on the situation. We will be eliciting diagrams from three types of material: **videos** of SL discourse to note down for the stenographer situation, **texts** in written French to translate to SL for the translator situation, and **topics** to talk about in SL speeches for the author situation.

Discourse genre

An essential property of SL is iconicity, i.e. a resemblance between the form of the performed signs and their meaning. For example, the form of the sign meaning "cat" in LSF is an outward gesture on both sides of the mouth depicting a cat's whiskers. Iconicity can be observed in an even broader sense. SL making a relevant use of the signing space, relative topological and geographical relationships between entities are expressed via a direct projection of the relationship inside it. Actions of agents targeted at their objects/patients also map the agent–patient direction directly into space, between the two entities involved, previously anchored as points in the signing space.

The highest use of the feature, often called "role shift", is where all the articulations produced by the signer become potentially relevant, conveying most of the meaning in a directly iconic way. In its most extreme form, it borders the effect of miming. The frequency of this feature is high in story telling for example, a genre keen on visual effects and contrasts, involving a lot of enacted situations. On the contrary, neutral statements about timeless facts or involving non-animated entities engage fewer instances of role shifts.

The graphical nature of the diagrams allows to transcribe a lot of those spatial arrangements and visual effects. It is therefore important that we collect instances of either end of that continuum if we want to characterise how they are transcribed when they are present, and what subtitutes for them when they are not. We decided to elicit discourses of three genres: **stories** and fables for discourses where iconicity and body engagement is preferred, and general **definitions** for examples reducing the use of role shifts and increasing that of neutral forms. We also included **news items** as an intermediate genre, journalists delivering neutral, disengaged discourse by construction but still involving many animated agents in time-anchored sequences of events.

A short-sized text example is given below for each genre:

> Story: "*Once upon a time, in the mountains, there was a caterpillar named Zoé who was green with large yellow spots. She was very pretty and very tender, but also very sad as she thought of her parents. They had become butterflies, and left her alone on the ground.*"

> Definition: "*The heart is a body organ located inside the rib cage and ensuring the blood flow. It is necessary to sustain life. Its stop causes death. Its inverted cone shape and red colour due to the presence of blood formed a well-known symbol for life and love.*"

> News item: "*At least 525 people have been killed in Indonesia by a tsunami caused by an underwater earthquake on Monday, according to a new count published by the government on Wednesday.*"

Discourse length

A lot of information can fit in a connected diagram. However, the longer the discourse captured by a diagram, the more separations will likely be observed, if only to allow for turning pages. The ways and reasons for diagram splits are yet to be studied, but we can expect differences attributable to the discourse length. Besides, longer SL utterances tend to introduce more of the context, set and agents

first before developing the actions taking place in the established scene. And the way diagrams organise these features as discourse length grows is also of interest. We therefore wanted some control and distribution on the discourse length.
For a first corpus with limited means, it was not possible to reach hour-long productions, and we decided to limit tasks to a couple of minutes. Besides, we have observed in earlier work that the average duration for a SL piece of discourse by heart limiting disfluencies (e.g. backtracking or insertions of filler gestures) was less than 1 min (Filhol and Hadjadj, 2018), which suggests that a comfortable memory buffer for delivering speech without reading notes is beneath this value. So we chose to collect examples on either side of this relevant boundary: those resulting in signed productions of less than 30 s (**short**), and those exceeding 1.5 min (**long**). We also left open a category for **isolated** clauses, e.g. "*I take my child to the swimming pool every other Wednesday.*". This is to allow testing particular language constructs in isolation which we could be curious about, although only a few were included since out-of-context elicitation is not representative of concrete use cases.

Discourse entities
Diagrams often depict agents and discourse entities as symbols (icons, written words, etc.) in certain positions, linked with arrows or other connecting graphics. Depending on their number, the graph can grow in complexity, or find other strategies if it becomes too densely connected. To enable such analysis, we chose to collect a set of productions in which the **number of participating entities** by task would be more or less evenly distributed across the tasks. Short lengths of natural discourse typically do not exhibit high counts of acting entities, but including examples of single agents as well as two-, three- and multiple-character scenes should enable first comparisons.
Note that this can be done only when providing texts or videos for elicitation. In the case of a topic assignment, because the discourse content is left up to the author, the entity count will not be controlled (but if relevant, it can be counted afterwards).

Placements and displacements
Relative movements and geographic and topological relationships play a special part in SL, inducing a heavy use of the signing space. In a similar way to what we did for discourse entities, we therefore wanted to ensure a variety of cases in terms of **number of placements and movements** in the discourse. In video elicitations, they refer to classifier placements, relocations, etc. In text elicitations, they refer to semantic equivalent, e.g. clauses like "*the rat rushed to the lion*" counted as a movement.

3.3. Task distribution and elicitation material

At this point we wanted to build a set of tasks that covered all possible genres, lengths, etc. We also wanted to limit the time each participant would spend on diagrams and signing the result. A total time spent of 4 hours being already enough to ask for, we decided to keep the time load under this value, using an indicative duration of 15 min for a short task, 45 min for a long one, 5 min for an isolated clause (which they will only be asked once at most), and a 2 min overhead time per task.
Furthermore, we needed to include various language user profiles. Translators and language professionals (teachers, linguists, etc.) as well as more naïve but native users, deaf and hearing provided they all qualified as fluent signers should be included and separated across task sets as appropriate. For example, participants with insecure understanding of written text should be assigned stenographer and author tasks, not translations. We constituted five sets of tasks allowing for such distribution: A and B, reserved for participants fluent in both written and signed language like interpreters or children of deaf adults, C and D, to be assigned mostly to deaf participants with SL as primary language (no text to read), and E, composed with professional deaf translators in mind (quite a unique profile in France) to compensate for the lack of text-elicited deaf productions.
Besides, we were interested in eliciting a few tasks in more than one set, in parallel:

- using the same elicitation material between different participant profiles;
- using the same contents in different modalities, i.e. translated beforehand, one in text and the other in video.

However, only a few could be done not to reduce the variety of the elicitation tasks too much. We summarise the chosen distribution in table 2, w.r.t. elicitation type (video, text, topic), length (short, long, isolated), and genre (story, news item, definition). Isolated clauses were chosen separately, to elicit specific semantic/language constructs and were not classified with a genre.
With the distribution in table 2, we ensure that with as few as 3 informants on each task set, we would collect 138 diagrams in total, including:

- 51 representing stories, 45 news items, and 30 definitions;
- 57 elicited by video (stenography), 66 by text (translations), and 15 by topic (free productions);
- 90 representing short discourse, 36 long discourse, and 12 isolated clauses
- two disjoint comparable subsets of 12 diagrams each, created through translation and stenography in parallel, half of them short and the other half long.

To choose the elicitation material for each task, we selected a dozen of texts for each length–genre pair, with length being either short or long, and genre being story, news item or definition. In the set were included translations of SL videos that were already available to us[3], and that could fit

[3]For example, we included the story available on the regional language LIMSI atlas, which we had in French (`https://atlas.limsi.fr/?tab=Hexagone`; select "Paris" to see the text) and in LSF (`https://atlas.limsi.fr/?tab=LNT`; select "LSF" to view the video) in parallel.

Task set	SL video (stenography)			Text to translate			Free topic		Est. time
	Short	Long	Isol.	Short	Long	Isol.	Short	Long	
A	1(a), 1, 0			3(b), 3, 1	0, 1, 0	1			207 min
B	0, 1, 1			2, 1, 0	1, 1, 1				226 min
C	1, 1, 1	1(d), 1, 1(e)	1				0, 0, 1		216 min
D	3(ab), 2(c), 1	1, 0, 0	1				1, 1, 0	0, 0, 1	237 min
E				1, 1(c), 1	1(d), 1, 1(e)	1	1, 0, 0		216 min

Triplet counts, in order: stories, news items, definitions.
Letters in parentheses: parallel elicitations (each refers to one content, appearing twice in the table).

Table 2: Elicitation task distribution

in either of the six sets. For each we counted the number of discourse acting entities, placements and movements. Then we distributed them evenly in the table until we approached the time load limit, making sure SL material was first put in the stenography cells, and that at least one entry for each of the count variables was put in the table. This ensures a variety of entity, placement and movement counts in the discourse contents across the whole set (though not per other variable combination, which would increase the combinatorics too much).

4. Data collection

We have begun following the protocol specified in the section above with a first set of participants. At this point we have collected the diagrams of 12 participants, assigned to the following task sets: 4 to set A, 4 to set B, 1 to set D, 3 to set E. This means that we already have a corpus of over 100 diagrams with their SL video counterpart, including:

- 42 stories, 41 news items, and 20 definitions;
- 21 elicited by video, 84 by text, and 6 by free topic;
- 76 short entries, 27 long, and 8 isolated clauses.

Figures 4, 5 and 6 give examples of collected diagrams, the last one being elicited with the "heart" definition exemplified at the end of §3.2.. We will keep enroling more participants as volunteers will manifest. Especially, we are reaching out to deaf groups so that we can better balance the elicitation profiles (grow the numbers on task sets C and D in particular). We then intend to deposit the corpus online, for example on the Ortolang platform well suited for this purpose[4], a year after we reach the 300 diagram threshold. Meanwhile, we can already observe that they all made extensive use of 2d graphics, though still dotted with written words, especially proper names (the top three in fig. 5). Words that are not proper names come in variable proportions (from 0 to 29 in a full A4 page, with a very uneven distribution). Also, we notice what would qualify as a preference for logographic symbols, capturing meaning instead of form. For example, the sun was always drawn as an icon representing a sun, not what the body/hands should do to sign sun. Incidentally, this is an opposite property to every proposed writing system for SL, and may explain the difficulty for those to catch on.

[4]Ortolang allows subsequent public access and download and data versioning. `www.ortolang.fr`

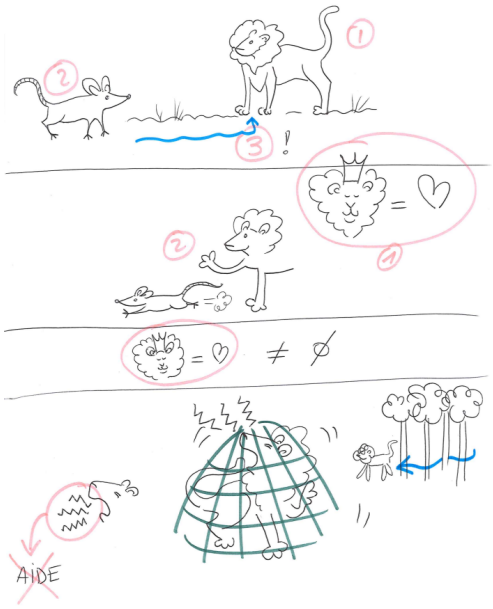


Figure 4: Diagram collected for a story

More specific patterns also appear to consistently link a graphical form and a meaning. These include:

- colour change (when available) for a focused event in a set up context, e.g. the blue arrows in fig. 4 denoting a path in an established scene;
- separation bars for the same reason, when the context is more abstract (an example was visible in fig. 3);
- the projection of topological constructions in the signing space onto the 2d plane, a feature already observed by Guitteny (2007) who studied SL discourse supported by educational (explanatory) diagrams;
- equal and comparison signs, as in figures 5 and 4;
- symbol repetition in enclosed shapes to mean sets of identical objects like in figure 5

More recurrent features can yet be found in the diagrams, the list above only being a sample. We explain these in an article to be published (Filhol, 2020), giving examples for

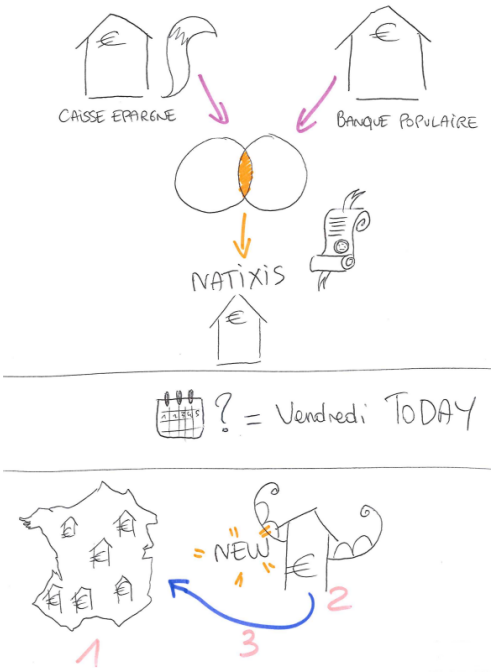


Figure 5: Diagram collected for a short news item

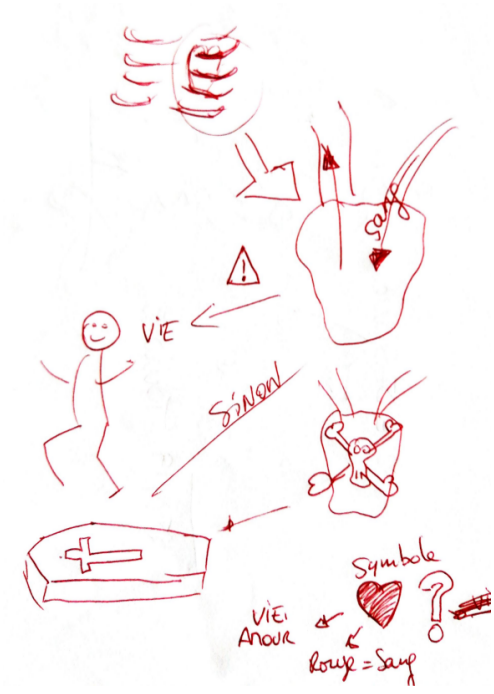


Figure 6: Diagram collected for a definition

each, and compare the general properties observed in the diagrams to those of writing systems and representations of SL.

5. Conclusion and prospects

After observing a few spontaneous SL representations by signers and having found that they exhibited recurrent features, we put forward that if patterns are found and studied in the productions, we would gain insight on the signers' approach to encoding their language. We then specified and applied an elicitation protocol to collect a corpus of such diagrams, in which major variables likely to impact their contents were balanced. We have begun data collection following the protocol and presented regularities which indicate that some underlying principles naturally come to the signers.

We propose that such regularities should not be ignored if so spontaneously produced by native and professional speakers of the language. They should instead seriously be investigated further as they may inspire some standardisation of SL representation that would be accepted by language users. Standardisation of such diagrams, if aimed at making them sharable and readable by other people than oneself to an arbitrary level of precision, may put the signing community on the track of shaping a new kind of writing system.

Acknowledgement

We wish to thank Interpretis[5] (Toulouse, France) for participating in the logistics of this corpus collection.

6. Bibliographical References

Athané, A. (2015). La schématisation : un travail original de préparation à la traduction de textes vers la langue des signes française. *Double Sens*, 4, Dec.

Barros, M. E. (2008). *ELiS – Escrita das Línguas de Sinais: proposta teórica e verificação prática*. Ph.D. thesis, Universidade federal de Santa Catarina, Centro de comunicação e expressão, Florianópolis.

Elliott, R., Glauert, J. R. W., Jennings, V., and Kennaway, J. R. (2004). An overview of the sigml notation and sigml signing software system. In *Proceedings of the 4th Language Resources and Evaluation Conference (LREC)*, pages 98–104, Lisbon, Portugal.

Elliott, R., Glauert, J. R. W., Kennaway, J. R., Marshall, I., and Sáfár, E. (2008). Linguistic modelling and language processing technologies for avatar-based sign language presentation. *Universal access in the information society (UAIS)*, 6(4):375–391.

Filhol, M. and Hadjadj, M. N. (2018). Elicitation protocol and material for a corpus of long prepared monologues in sign language. In *Proceedings of the workshop on Representation and Processing of Sign Languages*, Miyazaki, Japan, May.

Filhol, M. (2020). A human-editable sign language representation inspired by spontaneous productions... and a writing system? *Sign Language Studies*, 21(1).

Grushkin, D. A. (2017). Writing signed languages: What for? what form? *American Annals of the Deaf*, 161(5):509–527, Winter. Gallaudet University Press.

Guitteny, P. (2007). Langue des signes et schémas. *Traitement automatique de la langue (TAL)*, 48.

[5] http://interpretis.fr

Hanke, T. (2004). Hamnosys—representing sign language data in language resources and language processing contexts. In O. Streiter & C. Vettori, editor, *Proceedings of the workshop on the Representation and Processing of Sign Languages*, pages 1–6. European Language Resources Association (ELRA).

Kato, M. (2008). A study of notation and sign writing systems for the deaf. *Intercultural Communication Studies*, 17(4):97–114.

Lessa-de Oliveira, A. S. C. (2012). Libras escrita: o desafio de representar uma língua tridimensional por um sistema escrta linear. *Revista virtual de estudos da linguagem (ReVEL)*, 10(19):150–184. ISSN 1678-8931.

McCarty, A. L. (2004). Notation systems for reading and writing sign language. *The Analysis of Verbal Behavior*, 20:129–134.

Pizzuto, E. A. and Pietrandrea, P. (2001). The notation of signed texts: Open questions and indications for further research. *Sign Language & Linguistics*, 4(1–2):29–45, January.

Prillwitz, S., Leven, R., Zienert, H., Hanke, T., and Henning, J. (1989). Hamnosys version 2.0, hamburg notation system for sign languages, an introductory guide. *International studies on Sign Language communication of the Deaf*, 5. Signum press, Hamburg.

Seleskovitch, D. and Lederer, M. (1985). Interpréter pour traduire. *L'Information Grammaticale*, 25:44–47.

Supalla, S. and Blackburn, L. (2003). Learning how to read and bypassing sound. *Odyssey*, 5(1).

Sutton, V. (2014). *Lessons in SignWriting*. The SignWriting Press, 4th edition. ISBN 978-0-914336-55-6.

Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 61–68
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

The Synthesis of Complex Shape Deployments in Sign Language

Michael Filhol[1,2], **John McDonald**[3]
[1] Université Paris–Saclay, France
[2] CNRS, LIMSI, Orsay, France
[3] DePaul University, Chicago, IL, USA
michael.filhol@limsi.fr, jmcdonald@cdm.depaul.edu

Abstract

Proform constructs such as classifier predicates and size and shape specifiers are essential elements of Sign Language, but have remained a challenge for synthesis due to their highly variable nature. In contrast to frozen signs, which may be pre-animated or recorded, their variability necessitates a new approach both to their linguistic description and to their synthesis in animation. Though the specification and animation of classifier predicates was covered in previous works, size and shape specifiers have to this date remained unaddressed. This paper presents an efficient method for linguistically describing such specifiers using a small number of rules that cover a large range of possible constructs. It continues to show that with a small number of services in a signing avatar, these descriptions can be synthesized in a natural way that captures the essential gestural actions while also including the subtleties of human motion that make the signing legible.

Keywords: Sign Language Synthesis, Proforms, Classifiers, Size and Shape Specifiers, Avatars

1. Introduction

One of the unique aspects of Sign Language (SL) is its ability to make use of the signing space to locate, link and depict discourse entities in a dynamic manner. They involve iconic projections of topological relationships in the signing space, and symbolic use of spatial anchors for semantic relationships between them (Liddell, 2003; Johnston and Schembri, 2007).

The literature on such productive geometric (spatial) constructions often distinguishes at least two major types (Schembri, 2003; Woll, 2007; Zwitserlood, 2012):

- **classifier predicates**, involving language-specific handshapes (classifiers) that can be placed or moved to show relative positions and movements of semantically typed discourse entities (a person placed here, a car moving this way...);
- **size and shape specification**, involving language-constrained handshapes (SASSes) to describe the shape of an object and deploy lines or surfaces in space (e.g. the neck and body of a large vase).

They all have one purpose in common: description of relationships geometrically projected in the signing space, using handshapes conveying some semantic classification of what they stand for (person, vehicle, flat surface, small flat round object...). More than a handshape, they sometimes bring the whole arm into play (e.g. placing a tree), or a pair of handshapes for a single large object (e.g. a frame on wall). To avoid confusion in this paper, we will call all of these instances, whether used as a classifier or a SASS, a "**proform**". Proforms are mostly chosen from a language-specific list, e.g. (Vicars, 2020) for ASL, but one can observe others created on the fly.

In previous work, we covered a first set of SL constructions involving proforms (Filhol and McDonald, 2018; McDonald and Filhol, 2019) to demonstrate:

- the powerful geometric abstraction potential of AZee, a Sign Language modelling approach and description language;
- and the multi-track ability of Paula, a sign synthesis and animation rendering system.

Both designed to allow for parallel tracks and control of all body articulators, they have proven to be well-suited for one another (Filhol et al., 2017).

Among other concepts, we defined a *place-classifier* rule producing a small downward "settling" movement, whose meaning is to anchor an entity at a chosen point in the signing space, and a *move-classifier* rule producing a movement along a path in space, whose meaning is the displacement of the represented entity. This work mostly fell into the first type of proform use mentioned above: classifier predicates. The work reported in sections 2. and 3. of this paper addresses typical constructs of the second type, i.e. shape deployments, where proform movements outline shapes without meaning displacement. Then, in sections 4. through 6., it introduces new avatar animation techniques that facilitate the synthesis of these deployments while avoiding the robotic motion and unnatural postures seen in previous avatar synthesis systems.

2. Simple deployments

First, let us look at various examples of shape deployments[1]. Video "curtains" depicts two striped curtains hanging down from above the signer's face. There are two similar instances of deployments in the video, captured and annotated in figure 1. The first one delimits two sections alongside a window where the curtains are located, with a proform we shall call *thickness-medium*, useful to deploy strips of medium-sized breadth longitudinally. The second one draws stripes on each of these sections, with an eponymous proform *parallel-lines* that extends and spreads the fingers.

[1] Videos for quoted example names in this paper are available at `https://doi.org/10.5281/zenodo.3708057`.

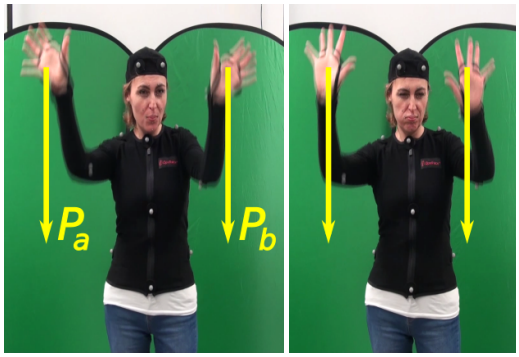


Figure 1: Two shape deployments in video “curtains”.

Figure 2: Two shape deployments in video “table”.

Video “table” starts by spreading a table surface with two flat hands (proform *flat-surface*) moving outwards from a point in the middle. This is immediately followed by a second deployment of the same table with a similar movement, only with a proform change to *thickness-medium*, which gives a second point of view on the table, this time as a delimited oblong shape. Annotated still images are given in figure 2.
Video “cupboard” is from the description of a piece of furniture against a wall. Its back is represented by the flat weak hand in the background (proform *flat-surface*), and a glass window in the front deployed on the strong hand in a vertical plane in the foreground (same proform). An annotated still shot of the relevant deployment construction is given in figure 3.
Continuing our prior work of formalising rules in AZee to represent more of the possible productions, we applied the AZee search methodology to a corpus containing many such examples (Benchiheub et al., 2016). This methodology is based on alternating searches for articulated forms and interpreted meanings in a corpus, and retrieving stable meaning–form associations. Branching and inverting each iteration by carrying over the common form/meaning counterpart as the starting point of the next, we establish production rules usable for synthesis (Hadjadj et al., 2018). Each production rule is a function that determines the forms to articulate for an identified meaning when applied (and given its arguments if any).

Figure 3: Shape deployment in video “cupboard”.

Looking at our examples, this method would lead to:

- interpret every annotated path in figures 1, 2 and 3 as **meaning** the depicted path (say P) deployed by the articulated proform (say *prf*);
- observe the same **form** (except for the differences accounted for by P and *prf*) for every instance, i.e. *prf* follows P with invariable dynamics, and the eye gaze is directed to the position of *prf*.

This observation warrants the definition of an AZee production rule, which we shall name *deploy-shape*, function of P and *prf*, carrying the meaning identified above and producing the form described above.
Say we now define a path $P_{L\to R}$ from left to right in the signing space. Applying the rule *deploy-shape* to it with, say, proform *flat-surface* as exemplified in E1 below would generate a sign score specifying a horizontal left-to-right sweep of the proform, then to be interpreted—although out of context here—as a flat surface deployed along the given path.

E1 *deploy-shape*($P_{L\to R}$, *flat-surface*)

All deployments in the proposed video example list apply *deploy-shape* on some level, each with its own argument values.
Examples “curtains” and “table” involve moments when two paths are deployed at the same time. Our earlier work introduced rule *simultaneous* to place two cups symmetrically on the table, or a knife and a fork on either side of a plate (McDonald and Filhol, 2019). Given two statements, the interpretation (meaning) of *simultaneous* is that both of them are true or happen at the same time, and its production (form) is the articulation of both simultaneously—typically on either side of the body. Shapes depicted by articulating two simultaneous deployments like in videos “curtains” and “table” verify the form produced by *simultaneous*, but also the corresponding interpretation: both parts of the shape exist at the same time, whether they are part of a same, physically continuous object (e.g. “table”) or not (e.g. “curtains”). The rule *simultaneous* is therefore well suited to capture this type of production, applying a *deploy-shape* on each hand.

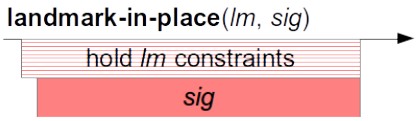


Figure 4: Box diagram for production rule *landmark-in-place*.

For example, given the two manual paths P_a and P_b shown in figure 1, the corresponding parallel deployments in video "curtains" can be represented by the AZee expression E2 below, with *prf=thickness-medium* for the first instance, and *prf=parallel-lines* for the second[2]. AZee terms "*s*" and "*w*" stand for strong and weak side respectively.

E2 *simultaneous*(*deploy-shape*(P_a, *prf*(hand=*s*)), *deploy-shape*(P_b, *prf*(hand=*w*)))

In video "cupboard", a hand is fixed, next to which a shape is deployed with the other hand. Sometimes, such examples make the fixed hand one side or end of the deployed shape, while the other takes care of the deployment from that point. Video "bedroom-walls" is an example of this, which we deal further with in the next section. In all of these instances, the fixed hand is interpreted as an active landmark in space, relative to which the rest is signed and potentially located as long as it is held in place. To capture this meaning–form association, we propose the new AZee production rule named *landmark-in-place*, function of a postural constraint lm for the held landmark and a signed piece of discourse sig, to be interpreted in the spatial context activated by lm. Its produced form is that of sig with lm installed just before sig starts, and held until sig ends (see fig. 4).
Given the location K of the background landmark and the path P_c shown in fig. 3, video "cupboard" can be accounted for by E3 below.

E3 *landmark-in-place*(*flat-surface*(hand=*w*, loc=K), *deploy-shape*(P_c, *flat-surface*(hand=*s*)))

3. Complex deployments

In this section, we take the challenge one level up and look at more complex shape deployments such as the one in video "bedroom-walls" (fig. 5). It describes the shape of a room by depicting two opposite walls, as illustrated in the diagram in fig. 6. The first wall (red) is made of two sections in a straight angle, while the second (blue) is made of one, reaching further out in distance from the front-most point. The video exhibits three horizontal manual strokes, one for each wall section (AB, BC and DE), plus an intermediate transitioning movement from C to D.
One immediately recognises shape deployments similar to those produced by *deploy-shape* defined above, and the use of *landmark-in-place* overarching the whole description in this case. To this extent the construct is comparable to E3, but the full utterance cannot be captured with a single followed path like in E3 because:

[2]Note that orientation issues are not specified in E2, for simplification. They will however be dealt with further down.

Figure 5: Shape deployment in video "bedroom-walls".

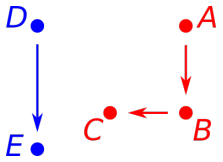


Figure 6: Layout of the wall sections in "bedroom-walls".

- the manual repositioning from C to D cannot be interpreted as a part of the depicted wall;
- the first two hand strokes do not follow a one-stroke dynamic that could be the result of a single followed path;
- the proform's orientation changes twice (at B and somewhere between C and D), without the rotation itself following a controlled path curve.

The utterance instead contains a combination of three straight deployments. Each one is describable by an expression similar to E1, but their combination remains an issue. They come one after the other, but following the AZee principle, no signed sequence is produced without identifiable production rules to justify it. In other words, the observed path concatenation must itself be the result of an application of defined production rules.
If we look at the signed forms visible in the video in more detail, we notice a few more clues:

- the first two strokes AB and BC are performed back to back, with shorter durations than the last or than those in the other videos;
- the last separate stroke DE is produced after a brief hold of the preceding posture (right hand at point C) and a quick but noticeable blink;
- another short hold is visible at position E, also with a blink.

These clues remind us of the forms produced by the rules *each-of* and *all-of*, also introduced in our work on classifiers (Filhol and McDonald, 2018). The rule *each-of* articulates an argument list of items in sequence, holding the final posture of each one for a short moment and appending a quick blink of the eyelids at the end of each hold. Its

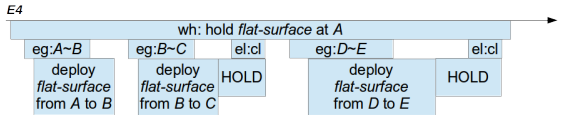


Figure 7: Sign score resulting from AZee expression E4.

meaning is that every element listed is equally true and focused, i.e. each is important, but none more salient than the other. It is a way in SL of building an exhaustive list of separate items, events or clauses of equal status. For example, our table scene description used it to lay out elements placed on a table.

The similar rule *all-of* also articulates an argument list of items in sequence, but each of them with accelerated dynamics (produced faster), and with no held feature in between. Its meaning also involves grouping the listed elements in a set of joint items or clauses, but this time focuses on the set as a whole, removing the relevance from its contained parts. It appeared in the table scene too, for example grouping 4 classifier placements to mean "[set of] 4 plates", and grouping signs KNIFE and FORK to mean "knife & fork pair"—cutlery items were then positioned pair by pair using two simultaneous classifiers for each pair.

By combining these two rules with *deploy-shape* and *landmark-in-place*, it is now possible to capture the form of "bedroom-walls" exactly, with expression E4 below[3]. The colours correspond to those in fig. 6, and are applied to the parts describing the matching wall sections.

E4 *landmark-in-place*(standing-wall(nrm=lat, hand=w, loc=A), *each-of*(*all-of*(*deploy-shape*($P_{A\to B}$, *standing-wall*(nrm=$-lat$)), *deploy-shape*($P_{B\to C}$, *standing-wall*(nrm=fwd))), *deploy-shape*($P_{D\to E}$, *standing-wall*(nrm=$-lat$))))

In this expression:

- *standing-wall* is a short-cut to specify *flat-surface* with the fingers up, leaving open the horizontal vector nrm which defines the plane in which the wall lies (normal orientation);
- $P_{X\to Y}$ is the straight path from point X to Y;
- lat an ipsilateral vector (pointing to the strong side), fwd a vector pointing forward (outwards from the body).

Expression E4 evaluates to the sign score (time line of signed forms) represented in figure 7, which matches the articulations visible in video "bedroom-walls" well. Moreover, like any AZee expression combining production rules, E4 not only produces forms but also conveys a composite meaning resulting from the semantic combination of the rules nested in the expression. In this case, E4 can be broken down into the following interpretation:

> *Drawing the scene from corner A of the room* [landmark-in-place]*, [there are] two separate wall sections* [each-of]*: one from A, made of two subsections* [all-of] *AB and BC, and the other from D to E.*

It appears that E4 gives the whole construct an interpretation that is entirely compatible with the meaning of the video.

4. Animating AZee

Given the above linguistic representations for sequences of deployments, the next task is to synthesize it as animation on a human avatar. The data provided by AZee consists of a series of timed blocks as in fig. 7, which are organized hierarchically. Each of these blocks controls different processes on the avatar's anatomy, which may all affect overlapping parts of the avatar's anatomy. For instance, the specifications for both the strong and weak hand "deploy flat surface" processes will affect the following parts of the anatomy (McDonald et al., 2017):

- the hands and arms to deploy the shape;
- the neck and eyes to direct gaze to the shape;
- the torso to support both of these processes.

The avatar must not only be able to schedule the sequence of required postures, it must be able to combine and blend their effects on the anatomy seamlessly.

More importantly, while each of these blocks contains animation information (e.g. as a sparse set of key body postures, or a mathematical procedure), this information is is necessarily an abstraction which defines just enough to carry the meaning, but which leaves out details of human motion that are essential to making signing look natural, legible and more human. This tug-of-war between the sparseness of the linguistic abstraction and the richness of human motion, has long plagued efforts to build avatars that synthesize sign directly from linguistic descriptions. However, the essential nature of this interplay has made it a key element of the bridge we have built between linguistics and animation, and has led to the present study.

Recall that the goal of the Paula sign synthesis system has been to leverage two key elements in an effort to animate sign legibly and naturally:

- the structure from a linguistic description of sign (Wolfe et al., 2015);
- the experienced eye and hand of an artist (Wolfe et al., 2011).

The previously proposed bridge (Filhol et al., 2017) from AZee sign descriptions to the animation of the Paula avatar is built on these two pillars through a system of templated shortcuts. These allow Paula to construct animations from larger blocks of motion, rather than the individual posture specifications that have driven prior avatars. These motion

[3] Unlike in previous examples, E4 includes orientation specifications, which change over the course of the depictions.

blocks can be of a range of types including mathematical procedures, pre-animated sequences, and hybrid procedures that draw significant posture and motion data from a small number of pre-animated sources.
Extending this bridge to animate the shape deployments described above becomes clear by reviewing the methods that the bridge has used for various types of discourse. When animating a frozen sign (whose form rarely changes other than for co-articulation or for ease of production) the system is free to shortcut to a pre-animated gestural unit that can be dropped in place of the AZee block and can then be blended with other elements.
These shortcuts were expanded to include templatable information in the prior case of classifier placement and movement (Filhol and McDonald, 2018). Here, the position of the placement and the direction of the movement change from production to production. However there are many other parameters that are left unconstrained by the linguistic description, and so may be set to whatever the animation system deems appropriate. This allowed Paula to leverage configuration data from an artist generated pose for the proform, which could include:

- the configuration of the hand for the proform;
- an orientation for the hand;
- the natural configuration of the shoulder and elbow;
- the accompanying configuration of the torso that supports the pose.

In this simpler situation, in comparison to the present study, the system was able to leverage this artist data because it is left unconstrained by AZee, i.e. not specified linguistically. When a parameter on the avatar is unconstrained, the system is free to choose a value for a parameter such as the height of the elbow, and wrist orientation that is comfortable and natural. It is precisely this comfort and naturalness provided by the artist that is one of the strengths of the Paula avatar. The additional fact that there are a limited number of commonly used classifiers, and the fact that the generic pose need only be set up once, makes this possible and not an undue burden on the artist.
Prior systems for generating sign movements directly from linguistic descriptions such as HamNoSys (Hanke, 2004) relied on automatic computations from inverse kinematics solutions, techniques originally designed to control robots (Buss, 2004), which contributed to the robotic nature of avatar motion from pure synthesis.
By leveraging an artist's eye for the pose and motion of the human body, the AZee-to-Paula bridge has been able to produce far more natural motion than prior systems were capable of. Another factor that contributes to the naturalness of motion generated by this system is the collection of ease controls for smooth motion control that are exploited as in (McDonald and Filhol, 2019).

5. Animating Deployments

The present study centers on a collection of shape deployments which constrain the system to a far greater degree than the prior classifier movements. These include:

Figure 8: Extreme wrist rotations in avatars

1. placement of objects, constrained in orientation that deviate significantly from the artist pose;
2. the deployment of surfaces in space described above, which will follow complicated orientations as the hand traces the surface shape.

Both of these situations will fully constrain the orientation of the hand in space relative to the body. For example, consider the deployment of a wall situated in front of the signer, and extending from left to right. The signer's palm will face the wall, i.e. out from the body, with fingers pointed up to show the wall's surface. The hand will then move toward the right to show the extension of the wall as laid out in figure 6, while maintaining the orientation.
These added constraints may at first seem like an advantage for the avatar, since it has far fewer unconstrained elements to fill. Unfortunately, from the perspective of producing natural motion it actually puts a straight-jacket on the avatar, forcing it into unnatural postures because of the coarseness of the linguistic specification as seen in the examples in figure 8 which are examples from the avatars described in (Kipp et al., 2011) and (Elliott et al., 2008). Such postures can even plague motion-capture derived signing due to the need for retargeting which often relies on the same kind of inverse kinematic solutions (Awad et al., 2009).
Paula would encounter the same issue in the second segment of expression E4, where the linguistic description specifies that the strong hand begin its motion at a medium distance in front of the weak shoulder with the hand facing forward and up against the horizontal wall. If the animation system were to attempt to orient the hand perfectly along these cardinal axes, the result would be seen in the left image in figure 9.
Of course, the human body never positions itself with such precision and considerations of comfort and strain will modify both the desired position and orientation. Notice that in figure 5, the hand is not pointed straight up nor is it facing perfectly forward. Yet, both the linguistic abstraction and the way that the resulting position is perceived by a viewer is consistent with an upward pointing-forward facing hand. The right image in figure 9 illustrates a more natural relaxed configuration of the hand.
The Paula system has, to this point, avoided this problem

Figure 9: Literal vs. relaxed interpretation of E4

through its reliance on an artist's touch. Unfortunately, in the present application, the infinite collection of possible positions and orientations precludes using an artist defined pose. So, the system must fall back on a more traditional application of inverse kinematics that treats the hand as a block to be positioned and oriented in space, with the shoulder and elbow rotated to place it there. The wrist is then forced into the orientation needed to obey the constraint, even if that orientation would break a human wrist.

6. Relaxing wrist orientations

To avoid such unnatural wrist postures, the new system introduces a relaxation algorithm which balances the linguistically specified spatial orientation with the perceived strain that the hand would be under. The human wrist's comfortable range of motion is a good start for this and is roughly the following (Gates et al., 2016).

- -40° to 40° for wrist flexion/extension;
- -25° to 25° for wrist ulnar/radial deviation;
- -60° to 60° for wrist flexion/extension.

To relax the hand and allow a more comfortable posture modification the Paula system applies a penalty to wrist angles outside the comfort range of motion[4]. Outside of this range, the angle will increase at a slower rate than would be specified linguistically until it reaches the maximum rotation of the joint, see figure 10.
If $\pm v_0$ is the discomfort free range, $\pm v_1$ is the maximum range of a joint, and $v = \frac{2(v_1 - v_0)}{\pi}$ then this can be achieved with a relaxation of:

$$v_0 + V \cdot \arctan\left(\frac{x - v_0}{V}\right)$$

Applying such a relaxation penalty to each of the rotations at the wrist yields the more natural pose seen on the right of fig. 9. The hand still reads as facing the far wall, but the wrist is no longer strained.

[4]In our implementation, these angles are set somewhat smaller than the physical ranges ($\pm 30°$, $\pm 15°$, $\pm 50°$ respectively) since the skin of the avatar shows strain slightly earlier than a human wrist would.

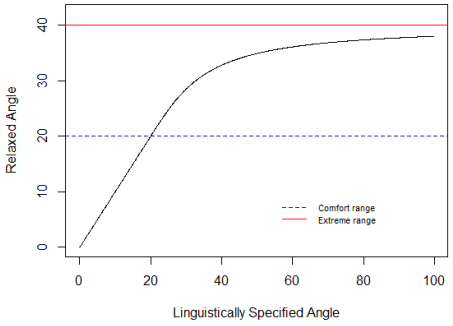


Figure 10: Angle Relaxation Function.

Figure 11: Synthesis of E2 with the Paula system.

7. Results

After implementing the features explained above on the Paula side of the system, the AZee parser was run on expressions E2 (*parallel-lines* in "curtains"), E3 ("cupboard") and E4 ("bedroom-walls"). The respective animations obtained can be found at `https://doi.org/10.5281/zenodo.3708057`, and illustrated in figures 11, 12 and 13.
The animations in these examples show that the avatar is able to follow the paths specified by the linguistic descrip-

Figure 12: Synthesis of E3 with the Paula system.

Figure 13: Synthesis of E4 with the Paula system.

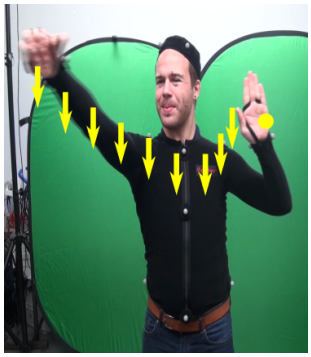


Figure 14: Repeated classifier placement with a landmark "in place".

tion, and further that the relaxation at work in these examples provides a naturalness that has previously not been achieved when driven directly from a linguistic description. Aside from the practical achievement of enlarging the set of AZee expressions that Paula is able to generate, we highlight that the type of structures rendered here, namely statements involving proforms, are of a non-fixed geometric kind that no other SL synthesis system has yet covered. By using geometric constructions as arguments of production rules, i.e. points, vectors... and transformations like scaling or translating, one can write an infinite number of AZee expressions, semantically composed and accounting for the ability in SL to make a productive, on-the-fly use of signing space.
Equally important is the fact that this expressive power emerges from a very small set of production rules. Indeed, the list of rules appearing in the reported AZee expressions in all of our works on proforms combined, in addition to the proforms themselves, are:

- *place-classifier*, producing the small "settle" movement ending at the point where to place the proform;
- *move-classifier*, making a proform follow a path and meaning to depict the displacement;
- *deploy-shape*, making a proform follow a path and meaning to depict the drawn shape;
- *simultaneous*, producing two statements at once and meaning that they are simultaneous;
- *landmark-in-place*, producing a statement while a fixed landmark is active;
- *each-of*, producing a list of separate items, each with equal importance;
- *all-of*, producing a set of items, with focus on the formed set.

In other words, the built-in geometric operators, a few proforms and seven production rules were enough to cover a large array of proform placements and shape deployments. This set includes productions whether they are used by themselves or in relation to one another, and whether they consist of a single stroke or of multiple paths.

Figure 15: Frozen sign RUSSIA while laying out a map with a landmark "in place".

Besides, in the AZee paradigm, whenever both a form and a meaning is found in a corpus to match those identified for a defined production rule, one can label the utterance as an application of the rule, provided its arguments can be identified as well. For example, we know that *landmark-in-place* can combine a given set of articulatory constraints (landmark argument *lm*) with any signing score (argument *sig*), which is to be interpreted in the spatial context of *lm*. We have seen this used to locate SASS deployments (e.g. "cupboard"), possibly repeated (e.g. "bedroom-walls"), but it can also be found with classifier placements or movements, also possibly repeated as exhibited in video "wine-bottles" (see fig. 14). What is more, it can combine with more complex scores mixing even dictionary signs like RUSSIA in example "map-layout", shown in figure 15.
Therefore, *landmark-in-place*, originally created for proforms in this paper, is not limited to proform constructions, let alone only to one type. Instead, it is much more generally applicable and transparently encompasses features that traditionally called for new concepts, like "buoys" (Liddell, 2003). Plus, *each-of*, *all-of* and *simultaneous* were created for expressions without proforms, and now used in this context.
In the light of this, we wish to emphasise the benefit of the general approach. Breaking down structures to arbitrarily deep levels and factoring elements into production

rules whenever consistent form–meaning associations are observed can provide insight on traditional linguistic categories.

8. Conclusion

This paper set out to extend the AZee coverage of Sign Language constructions depicting shapes deployed in space, as well as their animation with the Paula avatar. On the linguistic side, we introduced new rules such as *deploy-shape* and *landmark-in-place*, and reused prior rules like *simultaneous* when they fit the observed forms and carried the right meaning. On the synthesis side, we implemented new features such as geometric orientation of proforms and wrist relaxation to add naturalness to the postures where abstract linguistic specifications would otherwise lead to robotic or unnatural positions. With these efforts we managed to enlarge the set of proform constructions accounted for, from both of the AZee linguistic model and the Paula synthesis perspectives.

The naturalness in the output animation and expressive power in the input representation is encouraging and serves as an important validation step of both the linguistic model and the animation engine. Further, the synergy in the overall system drives forward the state of the art for both animation synthesis and linguistic representation, expanding the ability of avatars to produce even generative sign constructs such as proforms directly from linguistic descriptions.

Despite these significant gains in coverage, some aspects of proform constructions are still missing in our system, like proforms following curves, e.g. depicting a car taking a curve in the road. Future work will address these following a similar incremental methodology.

Bibliographical References

Awad, C., Courty, N., Duarte, K., Le Naour, T., and Gibet, S. (2009). A combined semantic and motion capture database for real-time sign language synthesis. In *International Workshop on Intelligent Virtual Agents*, pages 432–438. Springer.

Benchiheub, M., Berret, B., and Braffort, A. (2016). Collecting and analysing a motion-capture corpus of french sign language. In *7th International Conference on Language Resources and Evaluation (LREC), Workshop on the Representation and Processing of Sign Languages*, Portoroz, Slovenia.

Buss, S. R. (2004). Introduction to inverse kinematics with jacobian transpose, pseudoinverse and damped least squares methods. *IEEE Journal of Robotics and Automation*, 17(1-19):16.

Elliott, R., Glauert, J. R., Kennaway, J., Marshall, I., and Safar, E. (2008). Linguistic modelling and language-processing technologies for avatar-based sign language presentation. *Universal Access in the Information Society*, 6(4):375–391.

Filhol, M. and McDonald, J. (2018). Extending the AZee-Paula shortcuts to enable natural proform synthesis. In *Workshop on the Representation and Processing of Sign Languages*, Miyazaki, Japan, May.

Filhol, M., McDonald, J., and Wolfe, R. (2017). Synthesizing sign language by connecting linguistically structured descriptions to a multi-track animation system. *Universal Access in Human-Computer Interaction, Lecture Notes in Computer Science (Springer)*, 10278:27–40.

Gates, D. H., Walters, L. S., Cowley, J., Wilken, J. M., and Resnik, L. (2016). Range of motion requirements for upper-limb activities of daily living. *American Journal of Occupational Therapy*, 70(1):7001350010p1–7001350010p10.

Hadjadj, M., Filhol, M., and Braffort, A. (2018). Modeling French Sign Language: a proposal for a semantically compositional system. In ELRA, editor, *International Conference on Language Resources and Evaluation*, Miyazaki, Japan, May. ELRA.

Hanke, T. (2004). Hamnosys-representing sign language data in language resources and language processing contexts. In *LREC*, volume 4, pages 1–6.

Johnston, T. and Schembri, A. (2007). *Australian Sign Language (Auslan): an introduction to sign language linguistics*, volume 117. Cambridge, July.

Kipp, M., Heloir, A., and Nguyen, Q. (2011). Sign language avatars: Animation and comprehensibility. In *International Workshop on Intelligent Virtual Agents*, pages 113–126. Springer.

Liddell, S. (2003). *Grammar, gesture and meaning in American Sign Language*. Cambridge University Press.

McDonald, J. and Filhol, M. (2019). Fine Tuning Dynamics in Contextualized Proform Constructs from Linguistic Descriptions. In *International Workshop on Sign Language Translation and Avatar Technology*, Hamburg, Germany, September.

McDonald, J., Wolfe, R., Johnson, S., Baowidan, S., Moncrief, R., and Guo, N. (2017). An improved framework for layering linguistic processes in sign language generation: Why there should never be a "brows" tier. In *International Conference on Universal Access in Human-Computer Interaction*, pages 41–54. Springer.

Schembri, A., (2003). *Perspectives on Classifier Constructions in Sign Languages*, chapter Rethinking 'classifiers' in signed languages, pages 3–34. Psychology Press.

Vicars, W. (2020). ASL classifiers: `https://www.lifeprint.com/asl101/pages-signs/classifiers/classifiers-frame.htm`.

Wolfe, R., McDonald, J., and Schnepp, J. C. (2011). Avatar to depict sign language: Building from reusable hand animation. In *International Workshop on Sign Language Translation and Avatar Technology, Berlin, Germany*.

Wolfe, R., McDonald, J., Moncrief, R., Baowidan, S., and Stumbo, M. (2015). Inferring biomechanical kinematics from linguistic data: A case study for role shift. In *Symposium on Sign Language Translation and Avatar Technology (SLTAT), Paris, France*.

Woll, B. (2007). The linguistics of sign language classifiers: phonology, morpho-syntax, semantics and discourse. *Lingua: International Review of General Linguistics*, 117(7):1159–1353, July.

Zwitserlood, I., (2012). *Sign Language: an International Handbook*, chapter Classifiers, pages 158–186. Mouton de Gruyter, Berlin.

Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 69–74
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

Signing as Input for a Dictionary Query: Matching Signs Based on Joint Positions of the Dominant Hand

Manolis Fragkiadakis, Victoria Nyst, Peter van der Putten
Leiden University
Nonnensteeg 1-3 2311VJ, P.N. van Eyckhof 3 2311BV, Niels Bohrweg 1 2333CA
m.fragkiadakis@hum.leidenuniv.nl, v.a.s.nyst@hum.leidenuniv.nl, p.w.h.van.der.putten@liacs.leidenuniv.nl

Abstract

This study presents a new method to search sign language lexica, using a full sign as input for a query. Thus, a dictionary user can look up information about a sign by signing the sign to a webcam. The recorded sign is then compared to potential matching signs in the lexicon. As such, it provides a new way of searching sign language dictionaries to complement existing methods based on (spoken language) glosses or phonological features, like handshape or location. The "find the sign" method analyzes the recorded sign using OpenPose to extract the body and finger joint positions. To compare the recorded sign with the signs in the database, the variation in trajectories of the dominant hand and of the fingers is quantified and compared, using Dynamic Time Warping (DTW). The method was tested with ten people with various degrees of sign language proficiency. Each subject viewed a set of 20 out of 100 total signs from the newly compiled Ghanaian Sign Language lexicon and was asked to replicate the signs. The results show that our method can predict the matching sign with 87% and 74% accuracy at the Top-10 and Top-5 ranking level respectively by using only the trajectory of the dominant hand. Additionally, more proficient signers obtain 90% accuracy at the Top-10 ranking. The methodology has the potential to be used also as a variation measurement tool to quantify the difference in signing between different signers or sign languages in general.

Keywords: sign language lexica, search functionality, variation measurement

1. Introduction

In most sign language dictionaries, users can search a sign through a written gloss, a unique identifier that by definition refers to a sign. In some cases, the lexica offer the possibility to specify formal parameters of the target sign, for instance, its handshape and location (Figure 1). The Flemish Sign Language (VGT) dictionary (Van Herreweeghe et al., 2004), the Swedish Sign Language (Institutionen for Lingvistik, 2009) and the Danish Sign Language (Center for Tegnsprog, 2008) are some examples of such dictionaries. After the input, the user is offered a set of signs that match the selected properties which can be then viewed individually.

Although sign search functionality on the basis of a sign parameter value is a useful attribute of sign language lexica, dictionary compilers still have to link these values to the videos. Also, as Zwitserlood discusses, the users of such dictionaries must "abstract away from the sign as a whole" if they want to use the parameter search functionality (Zwitserlood, 2010). Even then, only signs that match the query 100% are returned, and there is no concept of an ordered set of results that match to some degree. A thorough overview of sign language lexica and their features can be found in Zwitserlood's review (2010).

In this paper we describe our "find the sign" methodology that allows inputting a full video-recorded sign to search for entries in a dictionary. This method requires no training of any kind of model such as the ones used for sign language recognition tasks. In its core, it is a comparison method to quantify the difference in the movement between signs. As a result, it can be used for any sign language. By utilizing a pre-trained pose estimation framework we extract the body and hand joint positions from users using their webcam. Subsequently, by employing Dynamic Time Warping we find the closest matching signs from a compiled lexicon. To date, this methodology has only been applied to sign language classification tasks (Jangyodsuk et al., 2014; Schneider et al., 2019; Ten Holt et al., 2007) and not as a mode to complement sign search possibly solving the problem of ordering retrieval previously discussed. Additionally, we have developed a visualization tool to allow researchers to view the rendered paths of the dominant hand to further explore the overall difference in signing movements.

The paper is structured as follows: in Section 2 we give an overview of methods that utilize Dynamic Time Warping in the gestural and sign language domain. In Section 3 we describe our methodology regarding the extraction of the body joint coordinates as well as the experimental setup, analysis, and visualization tool. In Section 4 we present the results of our experiments. We discuss them in Section 5 and conclude and motivate future research in Section 6.

Figure 1: Traditional search functionality as seen in the online Danish Sign Language dictionary (Center for Tegnsprog, 2008).

2. Related Work

Dynamic Time Warping (DTW) is a dynamic programming based time series comparison algorithm to produce a distance metric between two inputs. It has been widely used in the speech recognition domain since the early 1970's (Abdulla et al., 2003; Axelrod and Maison, 2004; Myers et al., 1980). While the original algorithm can be computationaly expensive, different variations have been developed over the years to reduce the overall complexity, with most notably the works of Itakura-Parallelogram (Itakura, 1975), Ratanamahatana-Koegh-Band (Ratanamahatana and Keogh, 2004) and Sakoe-Chiba-Band (Sakoe and Chiba, 2013).

As a technique, it has been long-established in the gesture and sign language recognition domain as well (Ahmed et al., 2016; Jambhale and Khaparde, 2014; Jangyodsuk et al., 2014). Due to the fact that it is a distance metric it requires no training and it is a perfect choice for applications where limited training samples are available.

Ten Holt and her colleagues presented an algorithm for Dynamic Time Warping (DTW) on multi-dimensional time series (MDDTW) to perform classification on 121 gestures recorded with two cameras in stereo position (Ten Holt et al., 2007). In Jangyodsuk et al. (2014) the authors investigated the use of DTW and Histogram of Oriented Gradient (HOG) to compare a query sign with those in a database of ASL signs using Kinect data. Their results showed an accuracy of 82% in a Top-10 ranking level.

Recent developments in the field of machine and deep learning have lead to advances in sign language and gesture recognition. However, these approaches pose restrictions to their overall applicability as they require large amount of data and computational power in order to be trained. Furthermore, proposed methods for sign language classfication have been based on special sensor hardware, such as Microsoft's Kinect presenting additional challenges in their duplicability as well difficulty in their technical set-up. Our proposed method does not require the use of depth data to extract the pose key-points as this is being held by the pretrained pose estimation framework OpenPose. This makes our approach suitable for any kind of sign language lexicon. Most recently, Schneider et al. (2019) used Dynamic Time Warping in conjuction with One-Nearest-Neighbor algorithm and OpenPose to perform classification on six gestures. Their results suggested an accuracy of 77.4%. A major advantage of their methodology is the necessity for very little training data. However, a considerable drawback of their study is that they have only tested a small amount of gestures. As a result, such as pipeline shows a major deterioration of the overall accuracy when an additional gesture is added into the classification task.

Our study repurposes the work of Schneider et al. by:

- considering signs instead of gestures as inputs in DTW
- extending significantly the number of signs used in the experiment
- adding the finger joints extracted by OpenPose as additional data
- testing whether signing proficiency infuences the accuracy of the method

3. Methodology

In this section we describe the pose estimation framework (i.e. OpenPose) as well as the apparatus and materials used in this study.

3.1. Pose Estimation

OpenPose is a real-time, open source for academic purposes library for multi-person 2D pose estimation (Cao et al., 2017). It can detect body, foot, hand and facial keypoints. It is a bottom-up approach meaning that it does not recognize first where a person is in an image and then extract the body joints but from the detection of the various key-points predicts the overall pose. In general, it exceeds in performance similar 2D body pose estimation libraries like Mask R-CNN (He et al., 2017) and Alpha-Pose (Li et al., 2018). Its major advantage lies in its high accuracy regardless of the number of people in an image or video.

OpenPose is able to run on different operating systems and hardware architectures while providing all the necessary tools for acquisition, visualization and output file generation. Its output consists of multiple json formatted files containing the pixel x, y coordinates of the body, hand and face joints. In this study only the body and hand predictions were used as the face joints were irrelevant for our purposes.

3.2. Preprocessing

The output of OpenPose consists of x,y pixel coordinates. As the people in each frame can potentially be in different locations, it is important to normalize their keypoints. Rotational invariance is omitted in this study as most people are expected to be in an upright position in front of the web camera. The normalization is done in two steps. Firstly, all the key points are translated in such way so that the neck key point shifts to the origo at (0,0). To accomplish the shift, the neck key points coordinates are substracted from all other key points. Secondly, the key points are scaled in such way so that the distance between the left and the right shoulder key point becomes 1. This is achieved by dividing all key points' coordinates by the distance between the left and right shoulder key point.The scale normalization method is based on previous studies by Celebi et al. (2013), Schneider et al. (2019) and Östling et al. (2018).

One additonal step added to the pipeline is the horizontal flip of the videos when a participant was left-handed. This step is achieved by measuring the average velocity of each hand. In cases where the left hand's velocity is greater than the respective of the right hand, a horizontal flip is applied. Such a process allows an independent handedness feature of the overall methodology.

3.3. Participants

Ten people were asked to participate in the research. Four of them have no experience with sign language whatsoever while the rest are experienced signers. Additionally, they were all informed about the general purpose of the research and gave their consent to participate. This study was approved by the Faculty ethics committee.

Identifier	Sign Gloss
s1	ABOUT
s2	BED
s3	BOOK
s4	CAPTAIN
s5	DREAM
s6	EAT
s7	ELEPHANT
s8	HISTORY
s9	HOTEL
s10	IF
s11	LAPTOP
s12	LATER
s13	LUNCH
s14	MEET
s15	MIND
s16	NEAR
s17	NOSE
s18	OPEN
s19	TALK
s20	TRUE

Table 1: List of signs shown to the participants of our experiment

3.4. Data

Each participant viewed only once a selection of 20 signs from the newly compiled Ghanaian Sign Language lexicon (HANDS!Lab, 2020). While the overall lexicon has more than 1300 signs we selected randomly 100 of them to be used in our experiments due to time limitations. The order was randomized for each participant to avoid potential biases. A full list can be seen in Table 1. Each video had a 1000 by 580 pixel resolution at 30 frames per second and lasted approximately 5 ($\pm$2) seconds. Recordings were made with a Macbook Pro's webcam at 1280 by 720 pixel resolution and 30 frames per second.

We employ the soft DTW method by Cuturi and Blondel (2017) deployed by the tslearn python package (Tavenard et al., 2017) to perform DTW on the normalized trajectories of the dominant hand. Their work takes advantage of a smoothed formulation of DTW that computes the soft-minimum of all alignment costs. In a pilot test we observed that soft DTW performed better compared to other DTW variants, and was thus used in the rest of the experiment. Furthermore, a DTW variant created by Sakoe and Chiba (2013) used by the same python module was utilized to measure the distance of the trajectories of all finger coordinates.

Most signs in our lexicon are one-handed where the left hand is inert either by being "absent" or passively fixed at a location. In the two-handed signs, the left hand mostly copies the movement of the right hand. As a result, we employed DTW only on the dominant hand features as the left hand would either be less informative or equally informative.

Finally, the limited resolution of the output from OpenPose had an undesired effect producing sudden spikes in the signal. This attribute has been previously acknowledged by Schneider et al. (2019) and was present here too. The videos in the lexicon were blurry when the hand was moving fast making OpenPose to mispredict the proper joint locations between consecutive frames. As such, failed to create a smooth path. To compensate for this behavior we included two additional steps. Firstly, all the dominant hand's wrist x,y coordinates that had a confidence level lower than 0.3 were deleted. Additionally, we used a median filter with radius r = 3 for smoothing the remaining signal. Moreover, we noticed that due to the good lighting conditions in the GSL lexicon there was a mismatch on the body joint's coordinates predicted by OpenPose. The lighting conditons of the videos captured with the participants were of poor quality making it hard for the DTW algorithm to operate properly. To solve that problem we decided to include in the lexicon the data from a random participant every time we tested the methodology. This step seems to add the necessary noise in the database that is nevertheless similar to the noise in the participants' data. As a result, the data of each participant's sign was compared with 120 signs in our database (100 from the GSL lexicon and 20 from another random participant). The overall pipeline can be seen in Figure 2.

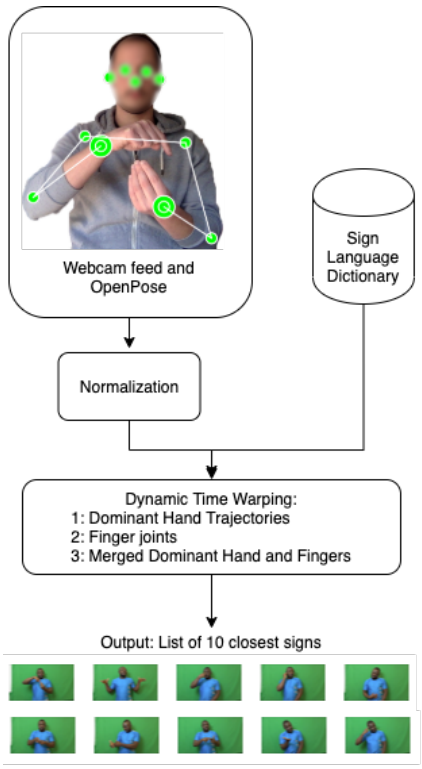
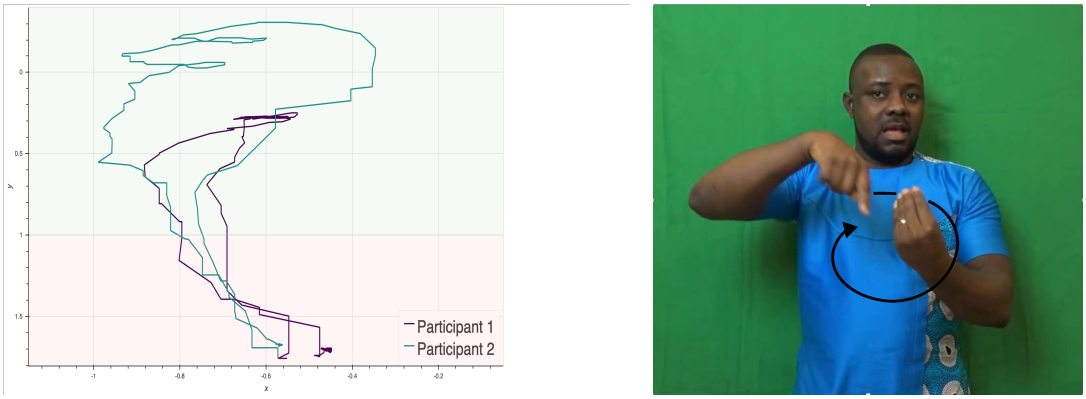


Figure 2: Overview of the overall pipeline of our methodology.

3.5. Visualization

To futher explore the outputs of OpenPose and how they are rendered in our methodology, we have created an interactive visualization tool. Developed with the python module "bokeh" (Bokeh Development Team, 2014), the user is able



(a) (b)

Figure 3: Visualization of the dominant hand trajectories between two participants (a) for the sign ABOUT (b).

to view the extracted dominant hand trajectories from the participants as a whole or individually. As all participants started and ended each sign in the same position, we have color coded as red the preparation and retraction phase and as green the stroke of each sign.

While the motivation behind the creation of this tool was to solely verify the output of openpose and the normalization part in our method, its potential reaches beyond the scope of this study. Such a tool, in combination with the DTW output, can potentially be used as a metric to quantify the variation in the movement and location of signers and sign languages in general. An example of the trajectories of two participants for the same sign can be seen in Figure 3a. It is evident that one participant produced the sign in a larger space with more distinctive movements. Moreover, it can be deduced that the location parameter is different as Participant 2 made the sign at a higher plane (almost in front of the face) while Participant 1 in front of the torso.

4. Results

Table 2 presents the overall accuracy of our methodology. Top-k refers to the number of signs a user must look up before finding a correct match. Accuracy indicates whether the target sign is present in the Top-k retrieved signs and is avereged accross all participants and signs.

It is evident that the highest accuracy is apparent at a Top-10 rank level at 87%. Furthermore, Top-5 rank shows an adequate accuracy at 74%. Contrary to expectations, using DTW in the joints of fingers extracted by OpenPose did not yield significant results with a highest accuracy at the Top-10 rank at approximatelly 52%. Merged DTW distances from the dominant hand trajectories and the finger joints also did not generate compelling results.

If only the experienced signers' data is considered then the accuracy at the Top-10 rank raises at 90% and the Top-5 at 78% (Table 2 row 4). On the other hand, the accuracy on the non-experienced signers drops at 82% and 0.67% at the Top-10 and 5 rank respectively (Table 2 row 5). Moreover, DTW on the finger's trajectories shows a significant drop at the Top-10 rank between the experienced and non-experienced signers of approximately 22% (Table 2 column 7).

The most striking observation to emerge from the analysis was that four out of 20 signs were consistently recognized with almost 100% accuracy at the Top-1 level rank. These signs were: CAPTAIN, DREAM, ELEPHANT and OPEN. Such behavior is justified as these signs have large, distinctive movements and locations that are hard to misinterpreted by the DTW.

5. Discussion

In this study we have investigated the use of OpenPose and Dynamic Time Warping as a ranking pipeline to retrieve matching signs from a sign language dictionary. Our results demonstrated that such a task can be achieved with an adequate accuracy rate.

This is in good agreement with the results obtained by Jangyodsuk et al. (2014). Although the accuracy rate does not match the one from Schneider et al. (2019) we have tested a larger vocabulary and lexicon. Additionally, we are not aiming at classifying each sign but rather create a suggestion ranking system. As such, our results suggest that approximately 9 out of 10 times the matched sign will be present in the first 10 retrieved signs.

Moreover, the results have further strengthened our hypothesis that signing proficiency is an influencing factor for

Condition	Dominant hand trajectory			Fingers' trajectories			Merged trajectories		
Top-k	Top 1	Top 5	Top 10	Top 1	Top 5	Top 10	Top 1	Top 5	Top 10
Accuracy of all participants	0,27	0,74	0,87	0,23	0,40	0,52	0,29	0,55	0,71
Accuracy of exprerienced signers	0,29	0,78	0,90	0,30	0,47	0,61	0,39	0,64	0,79
Accuracy of non-experienced signers	0,25	0,67	0,82	0,12	0,30	0,39	0,14	0,42	0,59

Table 2: Sign retrieval accuracy. Top k refers to number of best matches.

classification efforts. Although our sample size was limited there was a significant drop in the accuracy rates between the experienced and non-experienced signers. The former, produced well structured signs matching more appropriately the ones from the lexicon, which made DTW perform in a more excellent matter.

Our research failed to account for the low values of accuracy on the finger joints. This was probably as a result of the low performance of OpenPose in accurately predicting the finger joints due to low lighting conditions in the videos. It was often the case that joint predictions would dissapear between frames or mis-predicted in wrong locations. Thus, caution must be excercised when OpenPose is being used for such trivial tasks.

6. Conclusion

To sum up, we have obtained satisfactory results demonstrating the use of OpenPose and Dynamic Time Warping for a new, sign-based search functionality in reduced sign language dictionaries. We showed that our "find the sign" methodology can be used as a suggestion tool for sign retrieval in a small lexicon by using only the trajectory of the dominant hand. Additionally, our research has highlighted the importance of considering the level of signing proficiency when it comes to classification tasks. The significance of this study lies on the fact that the methodology in question can be easily used in any kind of sign language lexicon, irrespective of its quality and language. Additionally, no prior training of any kind of model is required. As such, this approach, in combination with the developed visualization module, has the potential to be used also as a metric tool to quantify the variation between signers and overall languages.

Furhtermore, a number of things is left for future work; first and foremost, to investigate how extracted finger joints can be utilized more efficiently in the overall pipeline. Moreover, different variants of the original DTW algorithms need to be tested. Finally, we intend to evaluate the use of other pose estimation frameworks, such as PoseNet, to further enhance the web and mobile user-friendliness of the method used.

7. Acknowledgements

We would like to thank all the people who participated in the study, without whose help this work would have never been possible.

8. Bibliographical References

Abdulla, W., Chow, D., and Sin, G. (2003). Cross-words reference template for dtw-based speech recognition systems. In *TENCON 2003. Conference on Convergent Technologies for Asia-Pacific Region*, volume 4, pages 1576–1579, Bangalore, India. Allied Publishers Pvt. Ltd.

Ahmed, W., Chanda, K., and Mitra, S. (2016). Vision based hand gesture recognition using dynamic time warping for indian sign language. In *2016 International Conference on Information Science (ICIS)*, pages 120–125, Kochi, India. IEEE.

Axelrod, S. and Maison, B. (2004). Combination of hidden Markov models with dynamic time warping for speech recognition. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–173–6, Montreal, Que., Canada. IEEE.

Bokeh Development Team. (2014). Bokeh: Python library for interactive visualization. `http://bokeh.pydata.org`.

Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1302–1310, Honolulu, HI. IEEE.

Celebi, S., Aydin, A. S., Talha, T. T., and Tarik, A. (2013). Gesture recognition using skeleton data with weighted dynamic time warping. In *Proceedings of the International Conference on Computer Vision Theory and Applications*, pages 620–625, Barcelona, Spain. SciTePress - Science and and Technology Publications.

Center for Tegnsprog. (2008). Ordbog over Dansk Tegnsprog. `http://www.tegnsprog.dk/`.

Cuturi, M. and Blondel, M. (2017). Soft-dtw: a differentiable loss function for time-series. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 894–903. JMLR. org.

HANDS!Lab. (2020). Ghanaian Sign Language. `https://play.google.com/store/apps/details?id=com.ljsharp.gsldictionary&hl=es_US`.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969.

Institutionen for Lingvistik. (2009). Svenskt teckenspraksleksikon. `https://teckenspraksleksikon.su.se`.

Itakura, F. (1975). Minimum prediction residual principle applied to speech recognition. *IEEE Transactions on acoustics, speech, and signal processing*, 23(1):67–72.

Jambhale, S. S. and Khaparde, A. (2014). Gesture recognition using DTW & piecewise DTW. In *2014 International Conference on Electronics and Communication Systems (ICECS)*, pages 1–5, Coimbatore. IEEE.

Jangyodsuk, P., Conly, C., and Athitsos, V. (2014). Sign language recognition using dynamic time warping and hand shape distance based on histogram of oriented gradient features. In *Proceedings of the 7th International Conference on PErvasive Technologies Related to Assistive Environments - PETRA '14*, pages 1–6, Rhodes, Greece. ACM Press.

Li, J., Wang, C., Zhu, H., Mao, Y., Fang, H.-S., and Lu, C. (2018). Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. *arXiv preprint arXiv:1812.00324*.

Myers, C., Rabiner, L., and Rosenberg, A. (1980). Performance tradeoffs in dynamic time warping algorithms for isolated word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(6):623–635.

Ratanamahatana, C. A. and Keogh, E. (2004). Making time-series classification more accurate using learned constraints. In *Proceedings of the 2004 SIAM Interna-*

tional Conference on Data Mining, pages 11–22. Society for Industrial and Applied Mathematics.

Sakoe, H. and Chiba, S. (2013). Gesture recognition using skeleton data with weighted dynamic time warping. In *Proceedings of the International Conference on Computer Vision Theory and Applications*, pages 620–625, Barcelona, Spain. SciTePress - Science and and Technology Publications.

Schneider, P., Memmesheimer, R., Kramer, I., and Paulus, D. (2019). Gesture recognition in rgb videos using human body keypoints and dynamic time warping. *arXiv:1906.12171 [cs]*. arXiv: 1906.12171.

Tavenard, R., Faouzi, J., Vandewiele, G., Divo, F., Androz, G., Holtz, C., Payne, M., Yurchak, R., Russwurm, M., Kolar, K., and Woods, E. (2017). tslearn: A machine learning toolkit dedicated to time-series data. `https://github.com/rtavenar/tslearn`.

Ten Holt, G. A., Reinders, M. J., and Hendriks, E. (2007). Multi-dimensional dynamic time warping for gesture recognition. In *Thirteenth annual conference of the Advanced School for Computing and Imaging*, volume 300, page 1.

Van Herreweeghe, M., Slembrouck, S., and Vermeerbergen, M. (2004). Digitaal Vlaamse Gebarentaal-Nederlands/Nederlands-Vlaamse Gebarentaal woordenboek. `https://woordenboek.vlaamsegebarentaal.be`.

Zwitserlood, I. (2010). Sign language lexicography in the early 21st century and a recently published dictionary of sign language of the netherlands. *International Journal of Lexicography*, 23(4):443–476.

Östling, R., Börstell, C., and Courtaux, S. (2018). Visual iconicity across sign languages: Large-scale automated video analysis of iconic articulators and locations. *Frontiers in Psychology*, 9:725.

Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 75–82
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

Extending the Public DGS Corpus in Size and Depth

Thomas Hanke, Marc Schulder, Reiner Konrad, Elena Jahn
Institute of German Sign Language and Communication of the Deaf
University of Hamburg, Germany
{thomas.hanke, marc.schulder, reiner.konrad, elena.jahn}@uni-hamburg.de

Abstract

In 2018 the DGS-Korpus project published the first full release of the Public DGS Corpus. This event marked a change of focus for the project. While before most attention had been on increasing the size of the corpus, now an increase in its depth became the priority. New data formats were added, corpus annotation conventions were released and OpenPose pose information was published for all transcripts. The community and research portal websites of the corpus also received upgrades, including persistent identifiers, archival copies of previous releases and improvements to their usability on mobile devices. The research portal was enhanced even further, improving its transcript web viewer, adding a KWIC concordance view, introducing cross-references to other linguistic resources of DGS and making its entire interface available in German in addition to English. This article provides an overview of these changes, chronicling the evolution of the Public DGS Corpus from its first release in 2018, through its second release in 2019 until its third release in 2020.

Keywords: German Sign Language (DGS), Linguistic Resource, Corpus, Resource Extension

1. Introduction

For the past eleven years, the DGS-Korpus project (Prillwitz et al., 2008) has been building the ***DGS Corpus***, an annotated collection of dialogues between native signers of German Sign Language (DGS). Based on this corpus, two publicly accessible resources are created by the project:

1. The ***Public DGS Corpus***, a subset of the full project corpus, accessible via two formats:
   (a) ***MY DGS***[1], a community portal for the Deaf community and others interested in DGS, which offers video recordings of selected dialogues with optional German subtitles, and
   (b) ***MY DGS – annotated***[2], a research portal for the international scientific community, which offers an annotated corpus of DGS for linguistic research.
2. ***Digitales Wörterbuch der Deutschen Gebärdensprache (DW-DGS)***, the first corpus-based digital dictionary of DGS–German.

These resources are released and extended progressively throughout the life time of the project. For example, while the first preliminary version of the community portal was released in 2015, its first full release, as well as the first release of the research portal, happened in 2018 (Jahn et al., 2018). The first preliminary release of the *DW-DGS* is scheduled for 2020.

In this article we present how the *Public DGS Corpus* and its two portals have been extended in subsequent releases after 2018. This involves the addition of new data, corrections to the subtitles and annotations, as well as several new features, such as new data formats, body pose information, unique identifiers, cross-references to other resources, collocation views and more. Some of these changes affect both the research and community portal, while others are only of relevance to researchers and therefore limited to the research portal.

The remainder of the article is structured as follows: Section 2 briefly introduces the DGS-Korpus project, its corpus creation efforts and how it has published corpus data up until the first full release of the *Public DGS Corpus* in 2018. The remaining chapters then address the changes introduced in 2019 (Release 2) and 2020 (Release 3). Section 3 presents the different kinds of content that have been added or extended, while Section 4 describes the different data formats in which data can be accessed on the research portal. Section 5 concludes the article by providing an outlook on the future directions that the corpus will take.

2. The *DGS Corpus*

The DGS-Korpus project[3] is a long-term project of the Academy of Sciences and Humanities in Hamburg. It was started in 2009 and aims to build a reference corpus of German Sign Language (DGS), publish a subset of about 50 hours with annotations in both German and English and to compile a corpus-based dictionary DGS – German. From 2010 to 2012 data were collected from 330 informants at 12 different locations in Germany. The selection of informants was balanced for sex, age, and region. The informants were filmed in pairs and presented 20 different elicitation tasks, which cover a broad variety of discussion formats and topics with a focus on dialogue and natural signing (Nishio et al., 2010). For information on the studio set-up, see Hanke et al. (2010).

The footage of the *DGS Corpus* consists of over 1150 hours of recordings, containing about 560 hours of near-natural DGS signing. The project uses iLex[4], an annotation tool and lexical database that was designed as a multi-user application for annotation and lemmatisation of sign language data (Hanke, 2002; Hanke and Storz, 2008).

The basic annotation of these videos comprises translation into German, lemmatisation and annotation of

[1] http://meine-dgs.de
[2] http://ling.meine-dgs.de
[3] http://dgs-korpus.de/
[4] www.sign-lang.uni-hamburg.de/ilex/

mouthing/mouth gestures. The translations were carried out by professional interpreters, alignment of these texts and further annotation mainly by student assistants. Lemma revision (Konrad and Langer, 2009; König et al., 2010) and detailed annotation are concerned with quality assurance and differentiating between morpho-syntactic inflection, modification, and phonological variation as a basis for the lexicographic analysis and description of signs.

In the following we will concentrate on the *Public DGS Corpus*. For information on the development of the dictionary DGS–German see Langer et al. (2018) and Wähl et al. (2018). For a discussion on how to link corpus and dictionary see Müller et al. (2020).

2.1. The Public DGS Corpus: One Corpus, Two Portals

The *Public DGS Corpus* is a 50 hour subset of the *DGS Corpus* intended for public release. In order to address the different needs of varying user groups (Jahn et al., 2018) access is provided via two different portals. As the DGS-Korpus project follows an open-access policy, both portals are freely accessible without any registration.

The first portal, *MY DGS*, addresses those interested in DGS, the history, life and culture of the deaf community. It contains over 47 hours of videos selected from the core elicitation tasks "Free conversation", "Discussion", "Subject areas", "Experience reports", "Region of origin", and "Deaf events" with German translations as optional subtitles, plus 2.4 hours of jokes (without translation).

The other portal, *MY DGS – annotated*, aims at an international audience that is interested in DGS data to perform their own research. In addition to the recordings of *MY DGS* it also contains 1.7 hours of recordings covering the remaining research-oriented elicitation tasks. These are included to provide examples of the variety of tasks in the *DGS Corpus*. Only two tasks are not part of the *Public DGS Corpus*: "Sign names" (for reasons of anonymisation) and "Isolated items" (elicitation using word and/or picture prompts). The videos are annotated with lemmas, mouthings/mouth gestures and translations. The research portal makes the videos and their annotations accessible both through a variety of downloadable file formats and through the portal website itself (see Section 4.1).

2.2. Release history

Videos and annotations of the *Public DGS Corpus* are being released and extended progressively throughout the life time of the project. To begin with, a pre-release of *MY DGS* containing ten hours of recordings was published in December 2015. Throughout 2016 and 2017 further recordings were added and improvements to the website were implemented. In May 2018 the first full release of the *Public DGS Corpus* was published (Jahn et al., 2018). This involved a content update to *MY DGS* and the first release of *MY DGS – annotated*. It increased the number of recording hours to 45.5. In February 2019 the annotation conventions of the corpus were added (Konrad et al., 2018).

Release 2, which was timed to coincide with with the TISLR 13 conference in September 2019, reached the project's target goal of 50 hours of publicly accessible recordings with almost 49 hours of lemmatised videos and more than 373,800 tokens. Starting with this release and continuing with Release 3, the focus of the *Public DGS Corpus* was shifted from adding size to adding depth.

3. New Features

In the following we report changes introduced in Release 2 and 3, focusing on new features. In Section 4 we also discuss the data formats of the corpus, the selection of which has also grown over time.

3.1. Persistent Identifiers and Archival Copies

As new versions of the *Public DGS Corpus* are released, previous versions are moved to publicly accessible archive directories. This raises the challenge of allowing users to clearly and persistently identify the version of the resource, e. g. for the purpose of citation. If a scientific article were to cite the research portal only by its URL, readers following the link could not be certain that they were viewing the same version of the corpus that the research of the article was based on, which would in turn affect the reproducibility of the research.

Instead, it has become good practice to identify resources by a persistent identifier, such as a **digital object identifier (DOI)**. DOIs allow objects, such as specific versions of a dataset, to be uniquely identified. A given DOI should always point to one and the same object and each object should only ever have one DOI. Different versions of an object should have different DOIs. A DOI is bound to metadata about its object, such as a URL at which it can be found. When the URL of the object changes, the metadata is updated to reflect this change, while the DOI stays the same. The metadata of a DOI can also be used to provide a description of the resource, citation information, version information and to connect it to the DOIs of other versions of the same data or to related resources.

In Release 2 we introduce DOIs for each release of *MY DGS* and *MY DGS – annotated*. Apart from DOIs for the overall web portals, we also provide DOIs for each individual video on the community portal and for each transcript and each sign type on the research portal. As part of this step, DOIs were generated not only for new releases, but also for the original first release.

Of course, there are also cases in which one might wish to refer to a resource in general, rather than to a specific version, e. g. to refer to a video on *MY DGS* via DOI (to be protected from changing URLs) while at the same time profiting from possible future corrections to its subtitles. For such purposes, Release 3 introduces **Concept DOIs**, which are DOIs that always point to the latest release of an object. Concept DOIs are created for all objects for which version-specific DOIs were previously created.

An important remark regarding our understanding of persistence: The corpus releases provide **semantic persistence**, but do not guarantee **byte persistence**. Semantic persistence refers to the semantic information that a resource provides. In the case of the corpus portals, this covers information like its recordings, their transcripts, the type names and token-type structure, keyword index, the reported statistics on data collection regions and informants

or the cross-references to other resources introduced in this paper (see Section 3.4). Any change to these kinds of information, regardless of whether it is the addition of new recordings or the correction of a single annotation or translation, results in a new release and the archival of the previous version.

Byte persistence, on the other hand, also implies that every byte of the digital files of a resource remains unchanged. This is undesirable for our purposes. For example, the HTML code of a page may have to be changed to ensure that it is rendered on modern web browsers or to update the hyperlink to another resource. These changes do not affect the semantic content of our resource and refraining from applying them would eventually result in archived releases becoming unusable due to purely technical reasons.

There are also certain components of the portal websites that we do not consider to be part of the semantic content of our resource for the purpose of persistence. These include, for example, the legal information provided in the imprint and data privacy information, which might have to be updated to comply with legal requirements. We also expressly exclude the title page of the community portal, as it presents a regularly changing selection of content, such as seasonal greetings and topic-specific compilations of stories from the corpus, e. g. how deaf Germans experienced the fall of the Berlin Wall.

It should also be noted that the exact specifications for data persistence of the *Public DGS Corpus* were determined during the months following Release 1 and a few additions were applied to the research portal retroactively. These changes, such as the linking of the annotation conventions (Section 3.3) and adding a second interface language (Section 3.7) are presented in this paper. None of these changes affect the corpus data itself. The archival version of Release 1 therefore represents the state of the research portal in February 2019.

3.2. Pose Information

To allow the computational processing of signed dialogues, we provide explicit machine-readable information on the location of various body parts, such as hands, shoulders, nose, ears, individual finger joints etc. This information is generated automatically using the pose estimation tool OpenPose (Cao et al., 2019). Apart from a general body model, which identifies major keypoints, such as elbows, shoulders, wrists, hip joints, eyes, nose, or ears, OpenPose can also compute detailed models of the face and each hand (Simon et al., 2017). An example of the computed information can be seen in Figure 1.

In Release 2 we introduced 2-dimensional pose information for perspectives *A* and *B*, the two frontal recordings of the participants. Release 3 adds pose information for perspective *Total*, which shows a side-view of both participants as they face each other. While the perspective also shows the moderator sitting between the participants and facing the camera, we choose not to include them in the pose information, as the moderator is not part of the corpus annotation (apart from translations of moderator utterances to aid the general understanding of the flow of conversation).

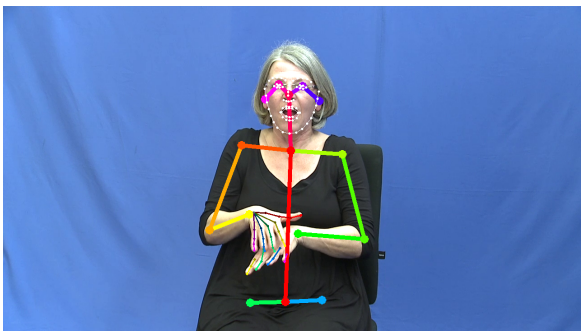


Figure 1: Visual representation of the pose information provided by OpenPose, computed for a video from the DGS-Korpus project. Sets of keypoints are generated for the body, the face and each hand. Lines between the points are added to the visual representation to indicate the logical connection between individual keypoints.

3.3. Annotation Conventions

The annotation conventions[5] were originally published as part of Release 1, explaining our approach of using a type hierarchy (double glossing) and double-token tags in iLex and the glossing conventions in *MY DGS – annotated* (Konrad et al., 2018). They were updated for Release 2 to report a change to the type-subtype relation of lexicalised forms of signs based on a manual alphabet, initialisation, and cued speech. In Release 3 we added the concordance view of tokens in each type entry (see Section 3.5) and introduced a *Sign/Lexeme* tier for each hand. Consequently, the description of "double-token tags" was updated.

3.4. Cross-References

The types list of *MY DGS – annotated* is automatically generated and shows all types and subtypes of the public corpus. For each type and its subtypes all tokens are listed below their respective gloss. In case that a studio reproduction of the citation form of the sign is available, the video is displayed under the gloss name. Studio recordings made for the *DW-DGS* show the isolated sign in four perspectives. Videos from prior productions provide a single perspective. As more dictionary entries are produced for the *DW-DGS*, more videos will also be added to the type list.

Release 3 also adds cross-references to lexical resources, namely to the *DW-DGS* and the language-for-specific-purposes (LSP) dictionaries *GaLex* (Konrad et al., 2010), *GLex* (Konrad et al., 2007), and *SLex* (Hanke et al., 2003).[6] Apart from the general value of cross-referencing resources, this also helps to contrast the difference between the type entries of *MY DGS – annotated* and the full lexical entries of the *DW-DGS*. For an example of a type entry with a multi-perspective video and cross-references, see the entry of `SMOOTH-OR-SLICK1^`[7] in Figure 2 . For a more in-depth discussion of our cross-referencing efforts, see Müller et al. (2020).

[5] `https://doi.org/10.25592/uhhfdm.822`

[6] Note that these lexical resources are not available in English.

[7] `https://doi.org/10.25592/dgs.corpus-3.0-type-13082`

Figure 2: View of the top part of the type entry for SMOOTH-OR-SLICK1^ on the research portal website. The video image shows the citation form of the sign. Below it are buttons to change the video perspective. To the right of the video are cross references to lexical entry 193 of the *DW-DGS* and to entries in *GLex* and *GaLex*.

Figure 3: Concordance view for type WEIRD1^, showing the first four tokens of the type in their context.

3.5. KWIC Concordance

Type entries on *MY DGS – annotated* list tokens for each type under the type or subtype gloss. In Releases 1 and 2, each token was represented as its gloss name and relevant metadata (region, format, age group, and sex) like e. g. EMBARRASSING2 Stuttgart | dgskorpus_stu_13 | 31-45f (Release 2 removes the indication of the elicitation task that was included in Release 1 (cf. Jahn et al., 2018)). The metadata information is also a hyperlink to the occurrence of the token within the transcript.

Release 3 significantly extends the type entries by displaying each token in a keyword-in-context (KWIC) concordance. KWIC concordance is a well-known tool in corpus linguistics in which a list of tokens of the search item (e. g. a word form or any annotated information) is given with its immediate context (items before and after, i. e. left and right neighbours). The list is centred around the search item. In the case of our type entries the KWIC concordance displays the tokens of each type and subtype with up to three neighbours left and right of the searched item. The metadata is displayed above the KWIC concordance. Next to it the translation of the utterance to which the token belongs is provided to give additional context. The gloss name of the token is integrated directly into the KWIC concordance.

Figure 3 shows the concordance of the first four tokens listed for the type WEIRD1^[8]. The entry for the first token is headed by its metadata, Berlin | dgskorpus_ber_08 | 31-60f, and the English translation of the utterance it is part of, “But at that point I didn’t really know what ‘being gay’ really meant.”. The translation tag limits the range out of which the left and right neighbours of the target token are taken. That’s why some concordances show less than three neighbour tokens left or right.

In addition, the KWIC concordance specifies the hand(s) the signer uses. The uppermost row displays a gloss when the sign is executed with the right hand, the middle row is for the left hand. The bottom row shows simultaneously articulated mouthings or mouth gestures. In the case of a mouthing spreading across multiple tokens, the annotation is centred in relation to the respective tokens.

For two-handed signs, the left and right hand rows are merged. As can be seen in Figure 3, when the dominance in two-handed symmetrical signs can be identified, the gloss is aligned with the row representing the dominant hand. Otherwise it is centred between the two rows. Consequently, the online transcript now also contains separate “Lexeme/Sign” columns for each hand (see Section 4.1.1).

KWIC concordances usually have a built-in function to sort the lines by left or right neighbour in order to look for collocations. This sort function is also be implemented in Release 3. Except for the target type, glosses in the KWIC concordance are clickable – like in the online transcript view – to open the respective type entry of the types list. The concordance view is also implemented in the dictionary entries of the *DW-DGS* (cf. Müller et al., 2020).

3.6. Usability on mobile devices

Since their introduction, smartphones and tablets have become more and more popular, replacing traditional desktop computers in many areas of life. In 2016 the internet use via mobile device passed desktop use for the first time (Statcounter GlobalStats, 2016). Making websites mobile-friendly has become a priority for web design. This requires layouts compatible with various screen sizes and touch-compatible navigation.

Furthermore, advances in web standards, such as the introduction of HTML 5, enabled the design of websites that are almost indistinguishable from regular mobile applications. Modern mobile operating systems even allow users to store websites as de-facto apps, representing them as a dedicated app icon on the home screen and hiding the web browser interface.

The *Public DGS Corpus* portals have always been designed with mobile-friendliness (and accessibility) in mind. As of Release 3, both portals can now also be stored as mobile apps, making access to the corpus even more seamless.

[8] https://doi.org/10.25592/dgs.corpus-3.0-type-18560

3.7. Update to Interface Languages

While the *Public DGS Corpus* was published with annotations in both German and English, the user interfaces of its portals were originally monolingual. The community portal was provided in German, as it is the written language most accessible to native speakers of DGS. The research portal had an English user interface, the lingua franca of the international research community. The only exception to this was the transcript web viewer (see Section 4.1.1), which allowed users to switch between German and English transcriptions (Jahn et al., 2018).

However, a side-effect of these language choices was the unintended implication that community members should not also be interested in the linguistic information available only on the research portal. Thanks to user feedback we became aware of this issue and upgraded the research portal user interface to be fully available in both English and German. This upgrade was also retroactively applied to Release 1, as it did not change the corpus itself and was deemed an urgent correction that should not wait until Release 2.

4. Data Formats

The page *Transcripts* on the research portal provides a list of all the dialogue transcripts that are part of the *Public DGS Corpus*. For each transcript it provides metadata (age group, elicitation format, topics of conversation) and a variety of file formats in which to access the corpus data. There are different file formats for annotation data (Section 4.1), video data (Section 4.2), pose information (Section 4.3) and metadata (Section 4.4).

4.1. Annotation Data

To support a variety of linguistic tools, the corpus annotations are made available in a number of different formats. They can be accessed via a web viewer (Section 4.1.1) or downloaded as *iLex*, *ELAN* or *SRT* files (Sections 4.1.2 to 4.1.4, respectively). All formats contain the basic annotation, i. e. translations, type glosses, and mouthings/mouth gestures. The inclusion of additional information depends on the limitations of each format.

4.1.1. Web Viewer

The web viewer provides a fast and easy way to directly inspect the *Public DGS Corpus* data without having to download it. It is is reachable via the research portal *MY DGS – annotated*, by clicking on the name of a transcript. It was first described by Jahn et al. (2018).

In the web viewer, the two informants are presented side by side in a video at the top of the page. Beneath it, the transcript is shown in vertical form (time flowing from top to bottom). The header of the transcript provides its name and a list of covered topics. In the vertical transcript, three tiers per informant are displayed: a translation (in either German or English), the tier *Lexeme/Sign* showing the glosses (in German or English) and the *Mouth* tier that displays mouthings (in German) or "[MG]" for any mouth gestures of the informant. A last column displays utterances or actions of the moderator as far as they are of (potential) relevance for the conversation.

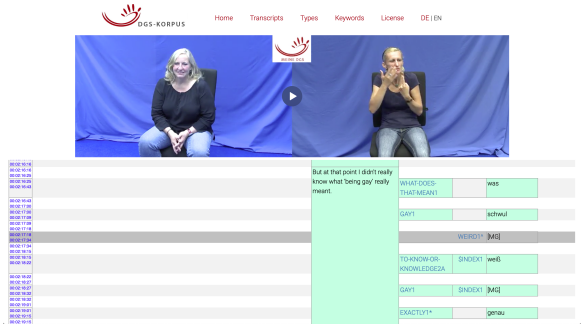


Figure 4: Web view of a transcript, with the video paused at a token of `WEIRD1^` articulated with the left hand being the active one.

Release 2 added DOIs for the respective transcript, placed above the transcript name. In Release 3, the *Translation* and *Mouth* tiers remain unchanged, but the *Sign/Lexeme* tier is now presented in two columns, one for the right and one for the left hand, as can be seen in Figure 4. This ensures consistency between the web viewer and the KWIC concordance view described in Section 3.5.

4.1.2. iLex

iLex (Hanke, 2002; Hanke and Storz, 2008) is an integrated annotation tool and lexical database, specifically designed to support consistent token-type matching (lemmatisation) and further annotation of sign language texts. The *DGS Corpus* and *Public DGS Corpus* were created using iLex, so it is naturally the tool that can model their information most accurately. Individual tokens refer to underlying type entries in the lexical database that are hierarchically structured into types and subtypes (Konrad et al., 2012; Konrad et al., 2018). The iLex files are the only available format that can explicitly represent the token-type relation and the type-subtype hierarchy.

Apart from a gloss, types also contain a phonetic transcription of their citation form using HamNoSys (Hanke, 2004). This should not be confused with a token transcription, which would also take into account deviations from the citation form. As iLex is the only one of our formats that can explicitly model the difference between tokens and types, and therefore the difference between token transcriptions and type transcriptions, we provide phonetic transcriptions only in this format.

Each transcript of the *Public DGS Corpus* is made available as an iLex *XML* file. Each file contains translations, gloss tokens, mouthings/mouth gesture annotations, gloss type hierarchies and phonetic HamNoSys transcriptions of types. When working with iLex the video recordings associated with an annotation can either be stored locally or accessed remotely on the *Public DGS Corpus* server.

4.1.3. ELAN

ELAN[9] (Crasborn and Sloetjes, 2008) is another popular annotation tool for sign language annotation. Information in ELAN is represented in tiers which are time-aligned to video files. The first time an .eaf file downloaded from *MY*

[9] `https://tla.mpi.nl/tools/tla-tools/elan/`

DGS – annotated is opened with ELAN, the location of the video files for the *A*, *B* and *Total* perspectives must be set. If files are not available locally, the users can also choose to work with fewer or none of the videos.
Each *ELAN* file provides 24 tiers that contain translations, lemmatisation and mouthings/mouth gestures. For each kind of information there exists a tier in English and in German. The only exception is the *Mouthing/Mouth Gesture* tier. Mouthings were not translated in English, as they refer to German words with different articulation features from e. g. mouthed English words. Within one language, one *Translation* tier each is provided for informant A, informant B and the moderator.
The lemmatisation by means of glosses is displayed in four tiers per informant and language. These result from the type hierarchy in iLex (two tiers for double glossing; see Konrad et al. (2018)) and the distribution to the active hand (two tiers for left and right hand):

Type hierarchy: In ELAN, type hierarchies are supported only indirectly. As ELAN does not support explicit type relations, each token is represented as the gloss of its type in a *Sign* tier. If the token has a subtype then the subtype gloss is represented in a *Lexeme/Sign* tier. For tokens connected directly to a type, the *Lexeme/Sign* tier is left empty.

Active hand: Depending on which hand is active in articulating the sign, tokens appear in tiers for either the right or left hand. For two-handed asymmetric signs the tiers of the active hand are filled. In the case of symmetric signs, the dominance of the hand is determined where possible, otherwise the right hand tiers are filled as default. Also, glosses for nonmanual activity like nonmanual gestures or exclusively oral activity are displayed in the right hand tiers (Konrad et al., 2018).

In ELAN, glosses represent tokens. The hierarchical relation between types and subtypes is lost. Therefore, information applied only to types but not tokens, such as the HamNoSys notation of citation forms, is not included in this format.

4.1.4. SRT

The SubRip Subtitle file format (*SRT*) is a popular format for storing subtitles separately from their video file. We provide our core annotation in this general-purpose format to allow its use with additional tools, such as MaxQDA[10], and in regular media players. In *SRT* files, text strings are associated with start and end timestamps to determine the time span in the video during which they should be displayed. It does not permit the inclusion of meta-information or the inclusion of multiple tracks to differentiate information. This means that there is no technical difference between type (glosses), mouthing, and translation items. To at least identify the origin of each utterance, each subtitle element starts with the identifying letter of its speaker (*A* and *B* for the participants, *C* for the moderator). The German and English data are provided in separate files.

[10] https://www.maxqda.com/

4.2. Video Data

All corpus recordings are provided as *MP4* video files, encoded using *H.264* compression at a resolution of 640 by 360 pixels and 50 frames per second.
Three perspectives are available: *Video A* and *Video B* each provide a frontal view of participant A and B, respectively. *Video Total* shows both participants from their side, facing each other, with the moderator sitting between them, facing the camera.
A fourth file, called *Video AB*, shows perspectives *A* and *B* next to each other. It corresponds to the video format shown in the web viewer (see Section 4.1.1) and on *MY DGS*. This file is provided for users that use the *SRT* format in applications that can only play a single video at a time.

4.3. Pose Information

The pose information for each transcript (see Section 3.2) is provided as a *JSON* file. To reduce its file size during transfer, the file is compressed using *gzip*. While OpenPose by default generates individual files for each frame, we compile all frames of all video perspectives in a single file. For users who require the default one-file-per-frame format, we provide a conversion script.[11]
Apart from the OpenPose output, the file also includes relevant metadata, such as the transcript ID, the camera perspective and the pixel dimensions of the original video on which OpenPose was run. Pixel dimensions are particularly important for users who wish to apply the pose information to the video files found on the research portal (see Section 4.2), as these are of smaller resolution than the original videos.
For further details on the OpenPose data of the *Public DGS Corpus* and its file format, please see the project note by Schulder and Hanke (2019).[12]

4.4. Metadata

Any kind of language resource will naturally have various kinds of metadata associated with it. This can be resource-wide information, like which language or languages the corpus contains, or information on specific parts, such as which age group individual informants belong to. To provide a standard for describing such language resource metadata, the Component MetaData Infrastructure (*CMDI*) was introduced in ISO 24622-1:2015 (2015).
The data formats we provide for the corpus have varying degrees of support for including metadata. To provide a single independent source for metadata, Release 3 introduces *CMDI XML* files for every transcript.

5. Outlook

One of the main motivations for many decisions regarding the design of the *Public DGS Corpus* and the changes made in its release versions was the feedback of users of the *Public DGS Corpus*. This shows for example in the changes the web viewer underwent throughout the releases or the addition of German as an interface language for *MY*

[11] https://github.com/DGS-Korpus/Public-Corpus-OpenPose-frame-extractor
[12] https://doi.org/10.25592/uhhfdm.842

DGS – annotated to not exclude user groups without English skills. While the web viewer of *MY DGS – annotated* was intended as a preview of the data that helps researchers select suitable data for download and further analysis with tools like iLex or ELAN, it turned out that many users prefer using the web viewer and expect to be able to do their research in it directly.

We look forward to feedback on new features such as the KWIC view. While we expect these features to mature over time and become sufficient for many purposes, this by no means replaces a full corpus research tool. As was announced in Jahn et al. (2018) we are also working on providing our data for ANNIS[13] (ANNotation of Information Structure).

6. Acknowledgements

We would like to thank the many student annotators who helped create the corpus and who contributed many of the corrections in its subsequent releases.

Also, we are very thankful for the valuable feedback provided by the community.

This publication has been produced in the context of the joint research funding of the German Federal Government and Federal States in the Academies' Programme, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies' Programme is coordinated by the Union of the Academies of Sciences and Humanities.

7. Bibliographical References

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., and Sheikh, Y. (2019). OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. arXiv preprint, v2.

Crasborn, O. and Sloetjes, H. (2008). Enhanced ELAN Functionality for Sign Language Corpora. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages at LREC*, pages 39–43, Marrakech, Morocco. European Language Resources Association.

Hanke, T. and Storz, J. (2008). iLex – A Database Tool for Integrating Sign Language Corpus Linguistics and Sign Language Lexicography. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages at LREC*, pages 64–67, Marrakech, Morocco. European Language Resources Association.

Hanke, T., Storz, J., and Wagner, S. (2010). iLex: Handling Multi-Camera Recordings. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages at LREC*, pages 110–111, Valletta, Malta. European Language Resources Association.

Hanke, T. (2002). iLex – A tool for Sign Language Lexicography and Corpus Analysis. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 923–926, Las Palmas, Canary Islands, Spain. European Language Resources Association.

Hanke, T. (2004). Hamnosys – Representing Sign Language Data in Language Resources and Language Processing Contexts. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages at LREC*, pages 1–6, Lisbon, Portugal. European Language Resources Association.

ISO 24622-1:2015. (2015). Language resource management — Component Metadata Infrastructure (CMDI) — Part 1: The Component Metadata Model. Standard, International Organization for Standardization, Geneva, Switzerland.

Jahn, E., Konrad, R., Langer, G., Wagner, S., and Hanke, T. (2018). Publishing DGS Corpus Data: Different Formats for Different Needs. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages at LREC*, pages 107–114, Miyazaki, Japan. European Language Resources Association.

König, S., Konrad, R., Langer, G., and Nishio, R. (2010). How Much Top-Down and Bottom-Up Do We Need to Build a Lemmatized Corpus? *Poster presented at the International Conference on Theoretical Issues in Sign Language Research*, West Lafayette, Indiana, USA.

Konrad, R. and Langer, G. (2009). Synergies between Transcription and Lexical Database Building: The Case of German Sign Language (DGS). In *Proceedings of the Corpus Linguistics Conference*, Liverpool, United Kingdom. University of Liverpool.

Konrad, R., Hanke, T., König, S., Langer, G., Matthes, S., Nishio, R., and Regen, A. (2012). From Form to Function. A Database Approach to Handle Lexicon Building and Spotting Token Forms in Sign Languages. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages at LREC*, pages 87–94, Istanbul, Turkey. European Language Resources Association.

Konrad, R., Hanke, T., Langer, G., König, S., König, L., Nishio, R., and Regen, A. (2018). Public DGS Corpus: Annotation Conventions. Project Note AP03-2018-01, DGS-Korpus project, IDGS, Hamburg University, Hamburg, Germany.

Langer, G., Müller, A., Wähl, S., and Bleicken, J. (2018). Authentic Examples in a Corpus-Based Sign Language Dictionary – Why and How. In *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, pages 483–497, Ljubljana, Slovenia. Ljubljana University Press.

Müller, A., Hanke, T., Konrad, R., Langer, G., and Wähl, S. (2020). From Dictionary to Corpus and Back Again – Linking Heterogeneous Language Resources for DGS. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages at LREC*, Marseille, France. European Language Resources Association.

Nishio, R., Hong, S.-E., König, S., Konrad, R., Langer, G., Hanke, T., and Rathmann, C. (2010). Elicitation Methods in the DGS (German Sign Language) Corpus Project. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages at LREC*, pages 178–185, Valletta, Malta. European Language Resources Association.

Prillwitz, S., Hanke, T., König, S., Konrad, R., Langer, G., and Schwarz, A. (2008). DGS Corpus Project – Development of a Corpus Based Electronic Dictionary German Sign Language / German. In *Proceedings of the Work-*

[13] http://corpus-tools.org/annis/

shop on the Representation and Processing of Sign Languages at LREC, pages 159–164, Marrakech, Morocco. European Language Resources Association.

Schulder, M. and Hanke, T. (2019). OpenPose in the Public DGS Corpus. Project Note AP06-2019-01, DGS-Korpus project, IDGS, Hamburg University, Hamburg, Germany.

Simon, T., Joo, H., Matthews, I., and Sheikh, Y. (2017). Hand Keypoint Detection in Single Images Using Multiview Bootstrapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4645–4653.

Statcounter GlobalStats. (2016). Mobile and tablet internet usage exceeds desktop for first time worldwide. https://gs.statcounter.com/press/mobile-and-tablet-internet-usage-exceeds-desktop-for-first-time-worldwide, November. Accessed: 2020-04-01.

Wähl, S., Langer, G., and Müller, A. (2018). Hand in Hand – Using Data from an Online Survey System to Support Lexicographic Work. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages at LREC*, pages 7–12, Miyazaki, Japan. European Language Resources Association.

8. Language Resource References

Hanke, T., Konrad, R., Schwarz, A., König, S., Langer, G., Pflugfelder, C., and Prillwitz, S. (2003). *Fachgebärdenlexikon Sozialarbeit/Sozialpädagogik*. Arbeitsgruppe Fachgebärdenlexika, IDGS, Hamburg University, URL http://www.sign-lang.uni-hamburg.de/slex/.

Hanke, T., König, S., Konrad, R., Langer, G., Barbeito Rey-Geißler, P., Blanck, D., Goldschmidt, S., Hofmann, I., Hong, S.-E., Jeziorski, O., Kleyboldt, T., König, L., Matthes, S., Nishio, R., Rathmann, C., Salden, U., Wagner, S., and Worseck, S. (2018). *MEINE DGS. Öffentliches Korpus der Deutschen Gebärdensprache, 1. Release*. DGS-Korpus project, IDGS, Hamburg University, DOI 10.25592/dgs.meinedgs-1.0.

Hanke, T., König, S., Konrad, R., Langer, G., Barbeito Rey-Geißler, P., Blanck, D., Goldschmidt, S., Hofmann, I., Hong, S.-E., Jeziorski, O., Kleyboldt, T., König, L., Matthes, S., Nishio, R., Rathmann, C., Salden, U., Wagner, S., and Worseck, S. (2019). *MEINE DGS. Öffentliches Korpus der Deutschen Gebärdensprache, 2. Release*. DGS-Korpus project, IDGS, Hamburg University, DOI 10.25592/dgs.meinedgs-2.0.

Hanke, T., König, S., Konrad, R., Langer, G., Barbeito Rey-Geißler, P., Blanck, D., Goldschmidt, S., Hofmann, I., Hong, S.-E., Jeziorski, O., Kleyboldt, T., König, L., Matthes, S., Nishio, R., Rathmann, C., Salden, U., Wagner, S., and Worseck, S. (2020). *MEINE DGS. Öffentliches Korpus der Deutschen Gebärdensprache, 3. Release*. DGS-Korpus project, IDGS, Hamburg University, DOI 10.25592/dgs.meinedgs-3.0.

Konrad, R., Langer, G., König, S., Hanke, T., and Prillwitz, S. (2007). *Fachgebärdenlexikon Gesundheit und Pflege*. Arbeitsgruppe Fachgebärdenlexika, IDGS, Hamburg University, URL http://www.sign-lang.uni-hamburg.de/glex/.

Konrad, R., Langer, G., König, S., Hanke, T., and Rathmann, C. (2010). *Fachgebärdenlexikon Gärtnerei und Landschaftsbau*. Arbeitsgruppe Fachgebärdenlexika, IDGS, Hamburg University, URL http://www.sign-lang.uni-hamburg.de/galex/.

Konrad, R., Hanke, T., Langer, G., Blanck, D., Bleicken, J., Hofmann, I., Jeziorski, O., König, L., König, S., Nishio, R., Regen, A., Salden, U., Wagner, S., and Worseck, S. (2018). *MY DGS – Annotated. Public Corpus of German Sign Language, 1st Release*. DGS-Korpus project, IDGS, Hamburg University, DOI 10.25592/dgs.corpus-1.0.

Konrad, R., Hanke, T., Langer, G., Blanck, D., Bleicken, J., Hofmann, I., Jeziorski, O., König, L., König, S., Nishio, R., Regen, A., Salden, U., Wagner, S., and Worseck, S. (2019). *MY DGS – Annotated. Public Corpus of German Sign Language, 2nd Release*. DGS-Korpus project, IDGS, Hamburg University, DOI 10.25592/dgs.corpus-2.0.

Konrad, R., Hanke, T., Langer, G., Blanck, D., Bleicken, J., Hofmann, I., Jeziorski, O., König, L., König, S., Nishio, R., Regen, A., Salden, U., Wagner, S., Worseck, S., and Schulder, M. (2020). *MY DGS – Annotated. Public Corpus of German Sign Language, 3rd Release*. DGS-Korpus project, IDGS, Hamburg University, DOI 10.25592/dgs.corpus-3.0.

Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 83–88
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

SignHunter – A Sign Elicitation Tool Suitable for Deaf Events

Thomas Hanke, Elena Jahn, Sabrina Wähl, Oliver Böse, Lutz König
Institute of German Sign Language and Communication of the Deaf
University of Hamburg, Germany
{thomas.hanke, elena.jahn, sabrina.waehl, oliver.boese, lutz.koenig}@uni-hamburg.de

Abstract

This paper presents *SignHunter*, a tool for collecting isolated signs, and discusses application possibilities. *SignHunter* is successfully used within the DGS-Korpus project to collect name signs for places and cities. The data adds to the content of a German Sign Language (DGS) – German dictionary which is currently being developed, as well as a freely accessible subset of the *DGS Corpus*, the *Public DGS Corpus*. We discuss reasons to complement a natural language corpus by eliciting concepts without context and present an application example of *SignHunter*.

Keywords: Elicitation tools, Mixed methods, Sign names for places and cities

1. Introduction

The use of corpora as language resources in the scientific exploration of natural language is commonly considered as state-of-the-art. However, building corpora, signed or spoken, large enough for the proficient analysis of a specific research question, is very costly in time, effort and budget. Building reference corpora that mirror natural language is even more challenging, as these corpora need to be of considerable size. As the size of a corpus determines to what extent low-frequency concepts are included in the corpus, and as the size of a corpus is limited at cost, some low-frequency concepts (e. g. discipline-specific vocabulary, regional specifications, cultural characteristics or names of cities, places and locations) will naturally be missing in each corpus.

This becomes an issue especially in the context of corpus-based dictionary creation where the user expects the semantic 'neighborhood' of each entry to be included as well. Therefore, lexicographers sometimes need to complement corpus-based dictionary entries with other low-frequency concepts that are not included (in enough quantity) in the corpus. In order to base such entries on data and not on the lexicographer's language intuition, supplementary data collection is needed.

In Langer et al. (2016), we introduced a system to collect data from members of the language community via a web-based application, the *Feedback System* (Wähl et al., 2018). This paper introduces *SignHunter*, a tool to elicit isolated concepts in any sign language at community events. *SignHunter* is used in the DGS-Korpus project to enhance the *DGS Corpus*, as well as a corpus based dictionary DGS-German with less-frequent sign names for places and cities.

2. The DGS-Korpus Project

The DGS-Korpus project is a long term project of the Academy of Sciences and Humanities in Hamburg (Prillwitz et al., 2008). The project's aims are:

- to build a reference corpus of DGS and to publish a subset of this corpus to be freely accessible online,
- to compile and publish a dictionary DGS-German that is based on and linked with the *DGS Corpus*.

2.1. The DGS Corpus

The *DGS Corpus* is designed as a reference corpus that displays the natural everyday language of deaf persons in Germany and is composed of 560 hours of signed narrations and dialogues. Parts of it have been translated, others have been annotated in detail. The elicitation took place in different regions across Germany to cover regional variants. The corpus is balanced for the sex of the informants, four age groups and the regions in which recordings took place. For the elicitation of the corpus, the tasks were deliberately designed to cover a broad variety of topics with as little influence on the informants as possible (Nishio et al., 2010). A translated and annotated subset of 50 hours of the *DGS Corpus* is published as the *Public DGS Corpus* (Jahn et al., 2018). The *Public DGS Corpus* is a research resource for natural DGS that is freely accessible online via two different portals, *MY DGS* (Hanke et al., 2020) for the DGS community and *MY DGS – annotated* (Konrad et al., 2020) for the research community.

One of the most important underlying motives that affected decisions regarding the design of the publication formats of the corpus was that the resource built should account for the needs of different user groups: persons who use DGS as their main language, interpreters, students, teachers and researchers interested not only in linguistic research but also in the history, culture and sociology of deaf persons across Germany, as well as many others. [1] This motive remains a driving factor in the improvement and enhancement of the resources published by the DGS-Korpus project

2.2. The DGS-German Dictionary

The "*Digitales Wörterbuch der Deutschen Gebärdensprache*" (*DW-DGS*) [Digital Dictionary of German Sign Language] is being compiled on the basis of the *DGS Corpus* data. Its final version is to be published in 2023, with the first pre-release made available in 2020. Information given on the signs include variants, typical mouthings, sense definitions, German translational equivalents, exam-

[1] For a more detailed description of the selection process, the data contained in the corpus and choices regarding the design of the two different portals, see Jahn et al. (2018).

ple sentences taken directly from the corpus (Langer et al., 2018), synonyms, antonyms, information on regionality, collocations and compound-like structures (Langer et al., 2019), as well as signs with a similar form and related signs.
Currently only signs that can be attested in the *DGS Corpus* are included and described in the dictionary. As the corpus is relatively large, it can be assumed that many of the commonly used signs and their meanings can be found in the corpus. Still some low-frequency signs will not be attested for in the corpus.
However, it would be desirable to have them listed (e. g. name signs for cities) in the upcoming dictionary as the information may be handy and interesting for the future user. This is where additional data collection methods can be used as long as it is transparent to the user which information is not based on the corpus.

3. SignHunter

SignHunter is an app created by the DGS-Korpus project that enables the user to collect isolated signs, the semantics of which are easy to communicate on a computer screen. It presents the informants a set of concepts they may choose from. This is the main difference between *SignHunter* and typical word list elicitations: Informants are free to choose what items to answer and how many.
Typically, a word list task requires the informant to answer all items on the list. This procedure may result in heavily skewed data as with sign languages informants could spontaneously invent new signs or fingerspell terms as they feel the pressure to give answers when prompted.
On the other hand, the approach taken here shares limitations with word list tasks, namely that it is crucial that there is no doubt about what concept the informant has in mind when producing a sign. We share concerns about word lists as the only basis for a dictionary (Brien and Brennan, 1995; Johnston and Schembri, 1999), therefore data from *SignHunter* can only be a supplement to the corpus data that were collected for the DGS-Korpus project.
In *SignHunter*, the set of concepts of interest to the researchers may be presented as a word list, a word list combined with visual stimuli or any other graphical representation of a set of concepts, such as a map showing geographical entities.

3.1. Data collection with SignHunter

Users are seated in front of the computer. Having identified herself by providing an id number, the system presents a set of concepts to the user. Having chosen an item, the user is invited to contribute her sign(s) for the respective item by signing into the camera.
Optionally, the user can playback the recording and then choose to either delete or keep the recording. In the latter case, the recording is automatically annotated with the concepts the user herself had identified. Of course this annotation needs to be examined and verified manually by human annotators later, but the automatic link between item and answer eases the annotation process considerably.
Once the user has provided one or more signs for the concept chosen, she can choose another concept to sign or just quit. Thus the informant herself operates the recording session.
SignHunter does not collect metadata, only the informant's id – a running number provided by the data collector team. Any metadata needed for further analyses needs to be collected separately. If used in public events, obviously the tool is best used in contexts where minimal data on the informants are sufficient. Depending on the program used for further processing the recordings, extra metadata need to be manually linked with the respective informants' ids.

3.2. Elicitation Setup

While the concepts available are simply pairs of ids and text labels in the *SignHunter* database, the system allows most flexible presentations for both the selection of concept and the prompt pages for concepts chosen by the informant: For this purpose, *SignHunter* just displays HTML pages that can be customized in the data collection preparation phase as needed.

3.3. Technical Details

SignHunter is an app that runs on macOS and Windows desktop or laptop systems and, once installed, can be used both online (i. e. with access to a database server) or offline. Obviously, the computer needs to have a camera built-in or attached.
During the elicitation sessions, all media files collected with *SignHunter* are stored locally, however in order to connect the media files to their metadata (signer, concept, date) *SignHunter* needs to connect to a database. If the computer *SignHunter* is running on cannot connect to the internet, a database can be installed locally. *SignHunter* uses PostgreSQL as the database machine which is easy and quick to install locally. It is also possible to record more than one person at a time by using multiple computers. In that case, a network needs to be set up., e. g. by connecting two computers with a network cable or a network switch or wifi router for more than two computers.
By using a second screen connected to one of the data collection computers (or another machine also in the network), the system can be used to compute and display output graphics at regular intervals. This allows the team to communicate the data collection progress to the audience, potentially attracting more informants to take part. An example of this is shown in Section 4 in Figure 1.

3.4. Further Processing

Once the recordings are finished, the video files collected need to be transferred to the central repository used in the annotation environment. In addition, the recordings table (holding the ids of the informants, the signed concepts, the name of the computer where the movie was recorded, as well as the ids of the movie files) needs to be transferred from the first machine's database server to your central server. *SignHunter* data collected by the DGS-Korpus project is stored in the annotation tool and lexical database *iLex*[2], a multi-user application for annotation and lemmatisation of sign language data (Hanke, 2002; Hanke and

[2] https://www.sign-lang.uni-hamburg.de/ilex/

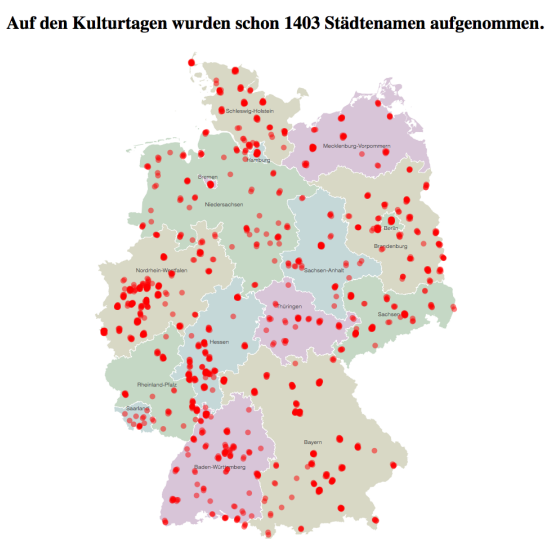


Figure 1: Count of recordings on the third/last day of the event. German caption: 'During the culture days, 1403 name signs for cities have already been collected'.

Storz, 2008). The fact that informants select the terms they want to sign and thus implicitly create an annotation of the target concept facilitates the annotation process to a great extent.

4. A Use Case: Name Signs for Cities

In addition to providing data to complement a corpus, collecting isolated signs with *SignHunter* is an excellent opportunity to directly engage a larger part of the language community beyond the group of 330 informants who participated in the corpus data collection. *SignHunter* made its first appearance at the the 6. Deutsche Kulturtage der Gehörlosen (6th German Culture Days of the Deaf) in 2018, a large event organized by and for the German Deaf Community. Name signs for cities and locations were collected with *SignHunter* on the three days of the event on two computers in parallel. A large screen showed the number of signs already collected during the event as well as dots for all cities on a map of Germany, as can be seen in Figure 1. The screen not only served to catch interest of bystanders, but was also an opportunity for team members to explain the data collection.

4.1. Elicitation Procedure

During the event, the name signs were collected by staff members of the DGS-Korpus project and student coworkers. The two computers were placed inside booths to guarantee good video quality without too much visual noise from bystanders and to allow some privacy. Informants were explained the goal and proceedings of the *SignHunter* elicitation and possible further uses of their data and had to read and sign an informed consent document. They were also explicitly told that they were free to chose what concepts and how many they want to record. Informants only needed to fill in their full name and address and sign

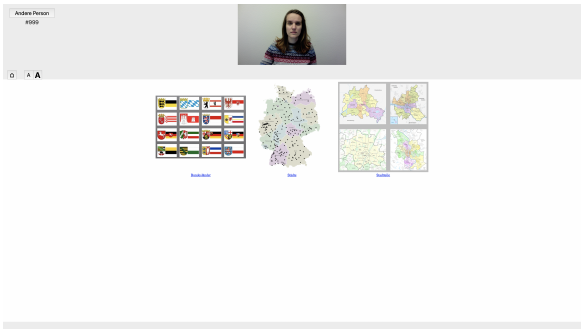


Figure 2: Selection of either federal states, German cities or districts of four large German cities (Berlin, Cologne, Hamburg, Munich).

the document. Further metadata like sex, age or others were not collected.

Each informant received a personalized id for logging in. The informed consent forms that were used for the elicitation were numbered, with the running number used as the informant ids. The concept chooser was a two-steps approach, in which the first step was for informants to choose between three options: whether they would like to record sign names for either federal states of Germany, German city names or the names for districts of four large German cities (Berlin, Cologne, Hamburg, Munich), as shown in Figure 2. Clicking on one of the options lead to a site on which the concepts were presented. In the case of German city names, items were presented both as a list of city names in German as well as a map of the state with the cities superimposed as dots (see Figure 3 The prompt page then contained the German name as well as some typical landmark, as exemplified in Figure 4.

The possible items for the cities were pre-selected, with cities with more than 100.000 inhabitants at the elicitation time being included in the item list as well as cities that have or had a Deaf school, Hard of Hearing school, a Deaf club or a Hard of Hearing Club.

In total, informants could choose from 470 items, out of which 363 were city names, 16 were names of federal states, 19 were districts of Munich and Cologne each, 26 were districts of Berlin and 27 were districts of Hamburg.

The recording was operated by the informants themselves. When finished with the recordings, the informant logged out. During the elicitation, a staff member or a student coworker was always present to answer (technical) questions. Due to the personalized id, it was possible for informants to return any time later and resume the recording.

4.2. Results

All in all, 135 persons took part in the *SignHunter* elicitation of name signs for German cities and locations. 404 distinct concepts were recorded. These 135 informants recorded 1978 answers in total, out of which 47 answers had to be excluded from analysis [3]. Answers were excluded for different reasons, mostly (in 35 cases) because the in-

[3]This as well as all following numbers are as evaluated in February 2020.

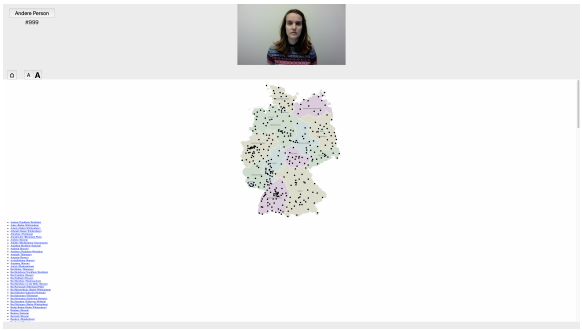


Figure 3: Selection of German cities either by dots on a map or by selecting city names from an alphabetical list.

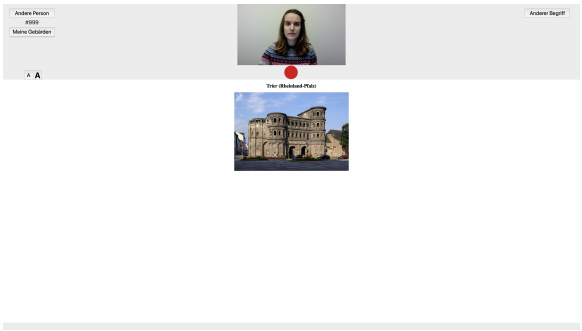


Figure 4: Example 'Trier'. German name of the city combined with a typical landmark as visual stimulus.

formant repeated the same answer twice or more. Another reason to exclude answers from further analysis was that the human annotators were unsure with respect to the assessment of the answer. From all 1978 recorded answers in total, in 47 cases the stimulus did not match the answer the informant gave. From these cases, 4 were excluded for other reasons from analysis, the other 43 cases were lemmatised nevertheless. In many mismatch cases, informants selected a federal state as stimulus and then recorded name signs for places in that respective federal state. In some cases, informants wanted to record sign names for places that were not included in the list of stimuli and thus used this little detour. As the answers recorded were nevertheless judged to be actual name signs and the informants seemed to have understood the task, these mismatch cases were not excluded from analysis. So, from the 1978 recorded answers, this leaves 1931 concepts recorded for lemmatisation and further analysis.

Out of these 1931 concepts, in 1558 cases the informants' answers were lemmatised with a single token. In 329 cases, there are two tokens per answer. This is predominantly due to the selected city names being compounds in German. In some cases the lemmatisation of two tokens per answer can also be traced back to reduplication. In a few cases, the answers were lemmatised into up to five tokens. For example, one informant signed every word of the German city name which is 'Brandenburg an der Havel'. The informant signed 'Brandenburg' as a compound-like structure [4] consisting of

[4] As these constructions are loans from German they are called compound-like here because they are not DGS compounds in the narrow sense (cf. Becker (2003)).

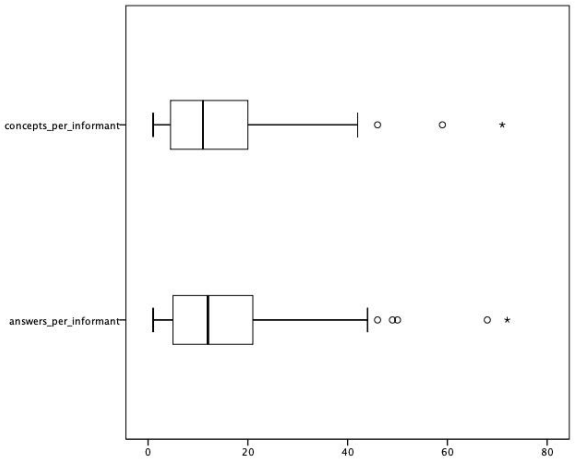


Figure 5: Number of recorded answers and recorded concepts per informant, with a total of 135 informants.

two signs and then added signs for 'an der Havel' [at the (river) Havel].

During the event, informants recorded between 1 and 72 answers, with the mean being 14.65 and the median being 12 answers per informant. The number of recorded concepts per informant ranges between 1 to 71, with the mean being 13.56 and the median being 11. As Figure 5 shows, both the total number of answers per informant and the recorded concepts per informant are distributed quite similarly, showing that informants recorded almost as many different concepts as answers. [5] The length of stay of informants in the recording cabins was not measured, but this data helps to predict the informants behavior for further elicitation at similar events.

As DGS is a sign language known to display a high number of variants (lexical as well as phonological) it could be expected that this might also show in the *SignHunter* recordings. However, the number of different variants attested for one concept was between 1 and 22, with the mean being 2.59 and the median being 2. As can be seen in Figure 6, for most concepts 1 to 7 variants were recorded, with more than 7 different variants per concepts being outliers in the data. The extreme outlier of 22 variants was attested for a concept that is a compound in German ("Baden-Württemberg") and was signed by most informants as a compound-like structure of which each part could have variants, e.g. for the first part "Baden" 3 possible variants were attested. This result mostly brings new challenges to light with respect to the representation of these units in the *DW-DGS*[6] as well as on *MY DGS* that need to be dealt with in the future.

The analysis of the collected data resulted in 123 new type

[5] For Figure 5 and Figure 6 the small circles represent outliers with a distance of 1.5 - 3.0*iqr (interquartile range) from the first or third quartile, stars represent extreme outliers with a distance of more than 3.0*iqr to the first or third quartile.

[6] Two questions arise here. One, how to represent the compound-like structure? Two, how to interlink them within the dictionary?

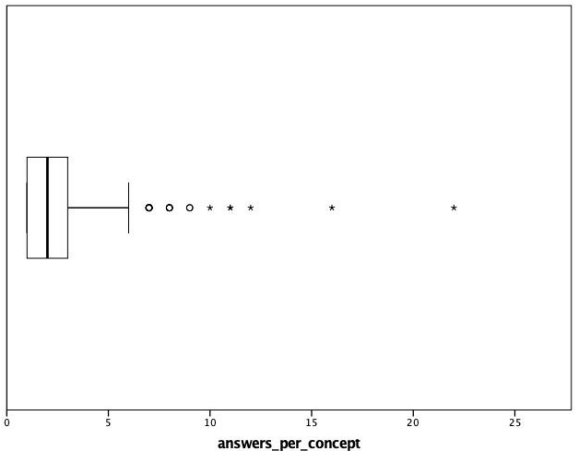


Figure 6: Number of recorded answers and recorded concepts per informant, with a total of 135 informants.

entries in our database. Thus our data could be supplemented and enriched by the data collection.

5. Conclusions and Outlook

The first application of *SignHunter* fulfilled our expectation for the data collection point of view, allowing us to include lists of city name signs in the dictionary on a solid data basis. In addition, the city name signs will be featured in the *MY DGS* community portal.

At the same time, feedback given during the event showed that many participants enjoyed the format. Setting up the HTML files for comparable cases is a straight-forward task. We include the project's focus group [7] into the decision process which vocabulary domains will be collected with *SignHunter* prospectively. By doing so we hope for future data collections to be relevant and interesting for the community as well as useful additions to the *DGS Corpus* and the *DW-DGS*. An interesting collection for future *SignHunter* recordings are for example name signs of famous deaf persons.

The focus group has also been trained in managing data collection sessions with *SignHunter*, allowing them to use the tool at events across Germany. As *SignHunter* runs on laptops, the equipment is relatively easy to transport and thus can be sent from one member of the focus group to another.

A feedback that was given during the event and that needs to be discussed for further elicitations is that some names informants would have liked to provide sign names for were not included in the list of concepts to choose from. In some cases, informants selected federal states as stimuli and then provided name signs for the concepts missing in the list. One possible solution would be to allow for free text input. However this would raise new difficulties, as the entered names may be ambiguous or may contain typos and thus would thus increase the annotation effort. Whether this is a useful addition should be checked by a test run first.

[7] The focus group is a group of deaf individuals actively taking part in the Deaf community. Members of the focus group are from different regions across Germany. The focus group cooperates with the DGS-Korpus project as advisers as well as multipliers, maintaining the contact with the DGS community.

6. Acknowledgments

We are thankful for the support of the student co-workers during the elicitation. We would also like to thank the members of the focus group who were engaged in the process of developing *SignHunter* and gave valuable feedback with respect to the user interface and the operation of *SignHunter*. This publication has been produced in the context of the joint research funding of the German Federal Government and Federal States in the Academies' Programme, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies' Programme is coordinated by the Union of the Academies of Sciences and Humanities.

7. Bibliographical References

Becker, C. (2003). *Verfahren der Lexikonerweiterung in der Deutschen Gebärdensprache*. Number 46 in International studies on sign language and the communication of the deaf. Signum, Seedorf, Germany.

Brien, D. and Brennan, M. (1995). Sign Language Dictionaries: Issues and Developments. In *Sign Language Research 1994: Proceedings of the Fourth European Congress on Sign Language Research*, pages 313–338, Munich, Germany. Hamburg : Signum.

Hanke, T. and Storz, J. (2008). iLex – A Database Tool for Integrating Sign Language Corpus Linguistics and Sign Language Lexicography. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages at LREC*, pages 64–67, Marrakech, Morocco. European Language Resources Association.

Hanke, T. (2002). iLex – A tool for Sign Language Lexicography and Corpus Analysis. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 923–926, Las Palmas, Canary Islands, Spain. European Language Resources Association.

Jahn, E., Konrad, R., Langer, G., Wagner, S., and Hanke, T. (2018). Publishing DGS Corpus Data: Different Formats for Different Needs. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages at LREC*, pages 107–114, Miyazaki, Japan. European Language Resources Association.

Johnston, T. and Schembri, A. C. (1999). On Defining Lexeme in a Signed language. *Sign Language & Linguistics*, 2(2):115–185.

Langer, G., König, S., Matthes, S., Groß, N., and Hanke, T. (2016). What Sign Language Lexicography Can Gain from a Mixed Method Approach: Corpus Data Supplemented by Crowd Sourcing. *Poster presented at the International Conference on Theoretical Issues in Sign Language Research*, Melbourne, Australia.

Langer, G., Müller, A., Wähl, S., and Bleicken, J. (2018). Authentic Examples in a Corpus-Based Sign Language Dictionary – Why and How. In *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, pages 483–497, Ljubljana, Slovenia. Ljubljana University Press.

Langer, G., Müller, A., Wähl, S., and Hanke, T. (2019). The DGS-Korpus approach to including frequent sign combinations in a corpus-based electronic sign language dictionary. *Poster presented at the International Conference on Theoretical Issues in Sign Language Research*, Hamburg, Germany.

Nishio, R., Hong, S.-E., König, S., Konrad, R., Langer, G., Hanke, T., and Rathmann, C. (2010). Elicitation Methods in the DGS (German Sign Language) Corpus Project. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages at LREC*, pages 178–185, Valletta, Malta. European Language Resources Association.

Prillwitz, S., Hanke, T., König, S., Konrad, R., Langer, G., and Schwarz, A. (2008). DGS Corpus Project – Development of a Corpus Based Electronic Dictionary German Sign Language / German. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages at LREC*, pages 159–164, Marrakech, Morocco. European Language Resources Association.

Wähl, S., Langer, G., and Müller, A. (2018). Hand in Hand – Using Data from an Online Survey System to Support Lexicographic Work. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages at LREC*, pages 7–12, Miyazaki, Japan. European Language Resources Association.

8. Language Resource References

Hanke, Thomas and König, Susanne and Konrad, Reiner and Langer, Gabriele and Barbeito Rey-Geißler, Patricia and Blanck, Dolly and Goldschmidt, Stefan and Hofmann, Ilona and Hong, Sung-Eun and Jeziorski, Olga and Kleyboldt, Thimo and König, Lutz and Matthes, Silke and Nishio, Rie and Rathmann, Christian and Salden, Uta and Wagner, Sven and Worseck, Satu. (2020). *MEINE DGS. Öffentliches Korpus der Deutschen Gebärdensprache, 3. Release*. DGS-Korpus project, IDGS, Hamburg University, DOI `10.25592/dgs.meinedgs-3.0`.

Konrad, Reiner and Hanke, Thomas and Langer, Gabriele and Blanck, Dolly and Bleicken, Julian and Hofmann, Ilona and Jeziorski, Olga and König, Lutz and König, Susanne and Nishio, Rie and Regen, Anja and Salden, Uta and Wagner, Sven and Worseck, Satu and Schulder, Marc. (2020). *MY DGS – Annotated. Public Corpus of German Sign Language, 3rd Release*. DGS-Korpus project, IDGS, Hamburg University, DOI `10.25592/dgs.corpus-3.0`.

Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 89–94
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

An Isolated-Signing RGBD Dataset of 100 American Sign Language Signs Produced by Fluent ASL Signers

Saad Hassan[1], Larwan Berke[1], Elahe Vahdani[2], Longlong Jing[2], Yingli Tian[2], Matt Huenerfauth[1]
[1] Rochester Institute of Technology, Rochester, NY 14623 USA
[2] The City College of New York, New York, NY 10031 USA
sh2513@rit.edu, lwb2627@rit.edu, evahdani@gradcenter.cuny.edu, ljing@gradcenter.cuny.edu, ytian@ccny.cuny.edu, matt.huenerfauth@rit.edu

Abstract
We have collected a new dataset consisting of color and depth videos of fluent American Sign Language (ASL) signers performing sequences of 100 ASL signs from a Kinect v2 sensor. This directed dataset had originally been collected as part of an ongoing collaborative project, to aid in the development of a sign-recognition system for identifying occurrences of these 100 signs in video. The set of words consist of vocabulary items that would commonly be learned in a first-year ASL course offered at a university, although the specific set of signs selected for inclusion in the dataset had been motivated by project-related factors. Given increasing interest among sign-recognition and other computer-vision researchers in red-green-blue-depth (RBGD) video, we release this dataset for use by the research community. In addition to the RGB video files, we share depth and HD face data as well as additional features of face, hands, and body produced through post-processing of this data.

Keywords: American Sign Language, dataset, RGBD video.

1. Introduction

Recently, progress in sensor technologies as well as research on algorithmic techniques supported by artificial intelligence methods has enabled the development of sign language recognition systems (Gkigkelos and Goumopoulos, 2017). Moreover, the availability of red-green-blue-depth (RGBD) sensors (Microsoft, 2014, 2020; Intel, 2020; Creative, 2013) has made it possible to capture depth maps in real time, facilitating many visual recognition tasks including ASL hand gesture recognition. Research on sign language and sign language recognition technologies can benefit from corpora that are collected using these RGBD cameras. This paper describes one such corpus that has been collected since April 2016 for facilitating research on sign recognition technology to be used for an educational tool.

We first describe the context and motivation of our work in section 2. In section 3, we summarize various existing datasets that are used to support research on sign languages and sign language recognition technologies. Section 4 describes the dataset in detail including the apparatus used, data collection methods, participant recruitment, and post-processing of the data. In section 5, we conclude with the insights we learned and some of the limitations of the dataset.

2. Motivation and Context

The release of our dataset is motivated by the increasing availability of RGBD video cameras as well as recent research on sign recognition that has considered RGBD video (discussed in section 3). As discussed below, some datasets have been collected to support various sign recognition research using a variety of camera systems, including the Intel RealSense, the Microsoft Kinect V2, and newer camera systems that have entered the market, e.g. (Microsoft, 2020). The low-cost of these consumer cameras has enabled the capture of high-resolution red-green-blue (RGB) videos with depth maps (D) (Ioannidou et al.,2017). These RGBD images provide photometric and geometric information not captured by traditional two-dimensional RGB camera systems.

While there have been various RGBD datasets collected in support of specific research projects, as discussed in section 3, the content of many of those datasets has been driven by the particular research interests of the particular team. Likewise, our new RGBD dataset was collected to support the development of a sign-recognition system, as part of a larger collaborative research project between City University of New York (CUNY) and Rochester Institute of Technology (RIT) (Huenerfauth et al., 2017; Ye et al., 2018; Huenerfauth et al., 2016) The computer vision team working at CUNY requested a targeted dataset to support the design of sign language recognition technology that would automatically analyze videos of ASL signing so that it can provide feedback to the user when particular errors are noticed in the video, e.g. as in the case of a student learning ASL who would like to practice their signing independently.

The specific goals of this larger project are not a focus of this paper, but we provide some details here to help explain the selection of the particular 100 ASL signs included in the dataset. As part of our ASL educational feedback system, a required sub-component is software that can identify occurrences of any of a set of 100 ASL signs that may appear during a video. These particular words were selected from among the vocabulary that is traditionally part of the first-year curriculum in most ASL courses offered at U.S. universities, and these particular words were selected since they related to some of the automatic error-detection rules that we intended to develop for our project. For instance, one rule determines whether the signer in the video has produced an ASL sign such as "NOT" that would typically require a negative headshake non-manual signal to be produced simultaneously. The system will indicate to the user that an error may have occurred if this manual sign is produced but a negative headshake was not performed. Additional details of our project appear in prior publications that describe human-computer interaction

research into the design of a system like this (Huenerfauth et al., 2017; Shah et al., 2019; Huenerfauth et al., 2016).

Thus, while we had originally collected this dataset for internal training purposes for creating one component of our research system (which explains the particular selection of the 100 ASL signs in this dataset), we decided to release this dataset for use by the community. Our decision has been motivated by an increased interest among computer vision researchers in working with color and depth data for human movement recognition. Thus, the ASL-100-RGBD dataset presented in this paper is disseminated for academic research on sign language recognition.

3. Existing ASL Databases

There are many publicly available corpora that have provided a valuable infrastructure for research on sign-language linguistics and for useful sign-related technologies for people who are deaf or hard-of-hearing. Traditionally, these videos consist of color video, but many were collected prior to the recent proliferation of RGBD video camera technology.

For instance, the National Center for Sign Language and Gesture Resources (NCSLGR) corpus contains ASL videos collected and linguistically annotated by researchers at Boston University. This dataset can be accessed using a web-based Data Access Interface (DAI), which provides access to data from the American Sign Language Linguistic Research Project (ASLLRP) (Neidle and Vogler, 2012; Neidle, 2002; Neidle, 2001). Several subsets of this database (Dreuw, Neidle, et al., 2008), including RWTH-BOSTON-50 and RWTH-BOSTON-104, were created in collaboration with RWTH Aachen University to build up benchmark databases for further research on sign language recognition. RWTH-BOSTON-50 was defined for assisting with the task of isolated sign language recognition (Zahedi et al., 2006). The RWTH-BOSTON-104 corpus has been used in continuous sign language recognition experiments (Dreuw et al., 2007; Dreuw, Stein, et al., 2008). Another commonly used sign language corpus of continuous signing data includes the RWTH-PHOENIX corpus consisting of German public TV station PHOENIX in the context of weather forecasts during daily news broadcast (Koller et al., 2015).

Similar to our new ASL-100-RGBD dataset, other ASL datasets consist of isolated sign productions. For instance, the American Sign Language Lexicon Video Dataset (ASLLVD) contains nearly 10,000 videos of over 3,300 ASL signs, produced by up to six native ASL signers in citation form, from multiple simultaneous camera angles, as well as various morphological and articulatory annotations for each (Athitsos, 2008). As another example, the Purdue RVL-SLLL ASL Database consists of 3576 videos from 14 ASL signers, and it was also collected using color video cameras, under two different lighting conditions (to suppress shadows or enhance contrast respectively). A portion of this corpus consists of continuous signing of memorized paragraphs, and another portion includes isolated sign productions (Martnez et al., 2002).

While these color-video corpora above (and many others beyond the few examples mentioned here) have been used in a variety of sign-language recognition research, there is emerging interest in the computer vision community at conducting research on data from sensors that provide both color and depth data, e.g. (Jing et al, 2019; Xie, 2018). More specifically, recent research has investigated sign recognition that considers a combination of RGB and depth information, e.g. (Almeidaab et al., 2014; Buehler et al., 2011, Chai et al., 2013; Jiang et al., 2014; Pugeault & Bowden, 2011; Ren et al., 2013, Yang, 2015; Ye et al., 2018; Zafrulla et al., 2011; Zhang et al., 2016).

Some of this research has considered static images with both color and depth information. For instance, Pugeault and Bowden investigated ASL fingerspelling letter recognition using a Kinect camera (2011). Keskin et al. captured data for 24 static images of handshapes as input to their classification model (2012). The American Sign Language Image Dataset (ASLID) contains 809 images (resolution 240 X 352) from various signs collected from six native ASL signers, as extracted from Gallaudet Dictionary videos. Ren et al. captured static handshapes for 10 ASL numerical digits using a Kinect camera, from 10 signers who were in visually cluttered backgrounds (2013).

The proliferation of RGBD video camera technology has propelled advances in areas such as reconstruction and gesture recognition. While the early RGBD data sets tended to be small (e.g. Bronstein et al., 2007), the field has expanded to include datasets for enabling research on identity recognition, pose recognition, and inferring facial expression and emotions (Min et al., 2014.; Fanelli et al., 2010; Firman, 2016). Recently these technological advancements have also enabled research on sign-recognition from RBGD videos. For instance, Yang developed a method to recognize 24 manual signs based on handshape and motion information extracted from RGBD videos (2015). Mehrotra et al. employed a support vector machine to recognize 37 Indian Sign Language (ISL) signs, based on 3D skeleton points captured using a Kinect Camera (2015). Kumar et al. used a combination of both a Leap Motion sensor and a Kinect Camera to recognize 50 ISL signs (2007). There has also been prior sign recognition research using RGBD video for Brazilian Sign Language (Almeidaab et al., 2014), Greek Sign Language (Gkigkelos and Goumopoulos, 2017), and Chinese Sign Language (USTC, 2019). Our dataset is also collected to exploit the depth modality for the recognition of strategically selected 100 ASL signs.

4. The ASL-100-RGBD Dataset

As discussed above, ASL-100-RGBD is a novel dataset that has been strategically collected and annotated to support the development of a sign language recognition system for use as a sub-component of our overall ASL education software system (Huenerfauth et al., 2017; Ye et al., 2018; Zhang et al., 2016). For that reason, the 100 ASL signs included in the dataset had been selected since they were signs commonly taught in the first-year curriculum of ASL courses in U.S. universities and because our system needed a detector for these ASL signs as part of some of its rules for providing feedback to users (Huenerfauth et al., 2017). Overall, the set of 100 signs includes some that are related to questions (e.g. WHERE, WHICH), negation (e.g. NONE, NOT), time-related words (e.g. TONIGHT,

TUESDAY). A full listing of the gloss labels used to identify these signs in our recordings is shown in Figure 3. The dataset consists of 100 ASL signs that have been produced by 22 fluent signers (details below), with each signer often producing multiple recordings. Each recorded video consists of the 100 ASL signs, and the start-time and end-time of each of the signs have been annotated, using the 100 text labels provided in Figure 3. Since this dataset had been collected for the internal development of a recognition system for our project, a custom set of gloss labels was used to identify each sign. The ASL-100-RGBD dataset is available via the Databrary platform (Huenerfauth, 2020). A sample video that visualizes the face and body-tracking information available in this dataset is available at the following URL: http://media-lab.ccny.cuny.edu/wordpress/datecode/.

4.1 Apparatus

The ASL-100-RGBD dataset has been captured by using a Kinect 2.0 RGBD camera. As shown in Figure 1, the output of this camera system includes multiple channels which include RGB, depth, skeleton joints (25 joints for every video frame), and HD face (1,347 points). The video resolution produced in 1920 x 1080 pixels for the RGB channel and 512 x 424 pixels for the depth channels respectively.

4.2 Data Collection

During the recording session, the participant was met by a member of our research team who was a native ASL signer. No other individuals were present during the data collection session. The participant was presented with a sequence of videos of a native ASL signer performing each of the desired 100 signs. Participants were asked to perform a sequence of the 100 individual ASL signs, without lowering their hands between signs. Signers were encouraged to hold their hands in a comfortable neutral position in the signing space in-between each of the signs. Time permitting, we collected two to three videos per signer, with each video containing up to one production of each of the 100 ASL signs. This process yielded a total collection of 42 video files, each containing about 100 signs and approximately 4,150 tokens in total.

4.3 Participants

All 22 of our participants were fluent ASL signers. As screening, we asked our participants: Did you use ASL at home growing up, or did you attend a school as a very young child where you used ASL? All the participants responded affirmatively to this question. A total of 22 DHH participants were recruited from the Rochester Institute of Technology campus. Participants included 15 men and 7 women, aged 20 to 51 (median = 23). Fifteen of our participants reported that they began using ASL when they were seven years old or younger. The remaining of the participants reported that they had been using ASL for at least 6 years and that they regularly used ASL at work or school.

4.4 Annotation and Post-Processing

The videos were annotated using ELAN, using the gloss labels shown in Figure 3, to indicate the start-time and stop-time of each token. At times, participants in our recordings accidentally omitted a requested sign, and at other times participants intentionally did not produce one of the requested signs. Participants in our video collection session were encouraged to produce a sign only if it were a sign that they would produce themselves; if they did not use a particular sign, e.g. due to some regional/dialectal variation, they were instructed to skip that sign. At other times in our videos, the participant accidentally performed a different sign than the specific form requested (as shown in the stimulus video). For this reason, our team needed to watch the resulting videos carefully to ensure that the signs included in the video were the specific 100 signs that had been requested. In the case of sign productions that differed from the designed token, e.g. with the signer using a different handshape or other variation, the sign was not annotated.

To make it easier for future researchers to make use of this dataset, we have also performed some post-processing of the Kinect data, with the output available as additional files in our dataset, accompanying each video. To extract the detailed coordinates of face, hands, and body from the RGB videos, we employed the OpenPose system (Cao et al., 2018), which is capable of detecting body, hand, facial, and foot keypoints of multiple people on single image in real time. The output of OpenPose includes estimation of 70 keypoints for the face including eyes, eyebrows, nose, mouth and face contour, e.g. as illustrated in Figure 2(a). The software also estimates 21 keypoints for each of the hands (Simon et al., 2017), including 3 keypoints for each finger, as shown in Figure 2(b). Additionally, there are 25 keypoints estimated for the body pose (and feet) (Cao et al., 2017; Wei et al., 2016), as shown in Figure 2(c).

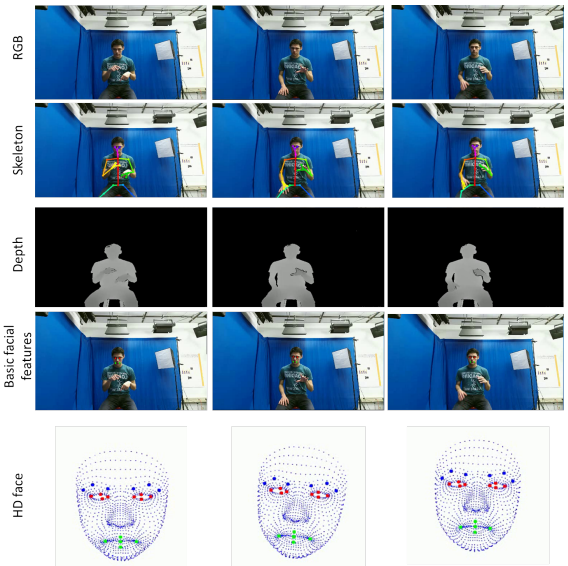


Figure 1: Samples of the available channels in our dataset including RGB, skeleton joints (25 joints for every frame), depth map, basic face features (5 main face components), and HD Face (1,347 points.)

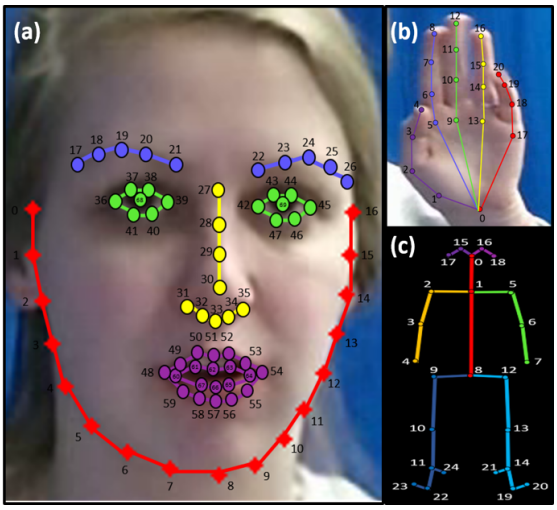


Figure 2: Figure 2: The coordinates of the extracted features from RGB channel for face, hand, and body by OpenPose.

```
ALWAYS, CAN'T_CANNOT, DODO1, DODO2,
DON'T_CARE, DON'T_KNOW, DON'T_LIKE,
DON'T_MIND, DON'T_WANT,
EIGHT_O_CLOCK1, EIGHT_O_CLOCK2,
ELEVEN_O_CLOCK, EVERY_AFTERNOON,
EVERY_DAY, EVERY_FRIDAY,
EVERY_MONDAY, EVERY_MORNING,
EVERY_NIGHT, EVERY_SATURDAY,
EVERY_SUNDAY, EVERY_THURSDAY,
EVERY_TUESDAY, EVERY_WEDNESDAY,
FIVE_O_CLOCK1, FIVE_O_CLOCK2,
FOR_FOR, FOUR_O_CLOCK1,
FOUR_O_CLOCK2, FRIDAY, HOW1, HOW2,
I_ME, IF_SUPPOSE, IX_HE_SHE_IT,
IX_THEY_THEM, LAST_WEEK, LAST_YEAR,
MIDNIGHT1, MONDAY, MONTH, MORNING,
NEVER, NEXT_WEEK1, NEXT_WEEK2,
NEXT_YEAR, NIGHT, NINE_O_CLOCK1,
NINE_O_CLOCK2, NO, NO_ONE, NONE,
NOON1, NOT, NOW, ONE_O_CLOCK1,
ONE_O_CLOCK2, PAST_PREVIOUS, QMWG,
QUESTION, RECENT, SATURDAY,
SEVEN_O_CLOCK1, SEVEN_O_CLOCK2,
SINCE_UP_TO_NOW, SIX_O_CLOCK1,
SIX_O_CLOCK2, SOMETIMES, SOON1,
SOON2, SUNDAY, TEN_O_CLOCK,
THREE_O_CLOCK1, THREE_O_CLOCK2,
THURSDAY, THURSDAY2, TIME, TODAY,
TOMORROW, TONIGHT, TUESDAY,
TWELVE_O_CLOCK, TWO_O_CLOCK1,
TWO_O_CLOCK2, WAVE_NO, WEDNESDAY,
WEEK, WHAT1, WHAT2, WHEN1, WHEN2,
WHERE, WHICH, WHO1, WHO2, WHO3,
WHY1, WHY2, WILL_FUTURE, YESTERDAY,
YOU
```

Figure 3. Gloss labels used in the ASL-100-RGBD dataset

5. Summary, Limitations, and Future Work

This paper has described the collection procedure and the contents of our new ASL-100-RGBD dataset. As described above, this dataset had originally been collected to support our project on designing an educational tool for providing feedback about potential errors during ASL signing, and we later decided to release this dataset for use by the research community.

Given its origins, there are several limitations of this dataset. For instance, the selection of 100 ASL signs in this dataset may seem somewhat arbitrary; the selection of this set had originally been driven by the specific needs of our research project. In addition, we have utilized a custom gloss label convention for labelling these signs (Figure 3), rather than aligning our gloss labeling with an established gloss convention used in prior ASL datasets. In addition, our dataset is small in size, and it only consists of data from 22 individuals, who primarily consist of young adults drawn from the Rochester Institute of Technology and surrounding community. For this reason, the individuals included in this dataset do not represent the wide variety of demographic and regional variation in ASL signing. Furthermore, the specific collection procedure used in this study employed a video stimulus presentation of an ASL sign performed by a native ASL signer. There is a risk that the artificial nature of this recording task could have influenced the naturalness of the ASL sign productions that were collected in this dataset.

In future work, we are utilizing this dataset to develop sign recognition software as part of our continuing efforts on our overall research project, which is focused on creating tools to provide feedback to ASL signers about potential errors in videos of their ASL signing.

6. Acknowledgements

This material is based upon work supported by the National Science Foundation under award Nos. 1462280, 1400802, and 1400810. We are grateful for the contributions of our collaborator Elaine Gale at CUNY Hunter College and for our research assistants Kasmira Patel and Anmolvir Kaur.

7. Bibliographical References

Almeidaab, S.G.M., Guimaresc, F.G., Ramrez, J. (2014). Feature extraction in Brazilian sign language recognition based on phonological structure and using RGB-D sensors. *Expert Systems with Applications* 41(16), 7259–7271

Athitsos, V., Neidle, C., Sclaroff, S., Nash, J., Stefan, A., Yuan, Q., Thangali, A. (2008). The ASL lexicon video dataset. In: *Proceedings of CVPR 2008 Workshop on Human Communicative Behaviour Analysis.* IEEE.

Bronstein, A. M., Bronstein, M. M., & Kimmel, R. (2007). Calculus of Nonrigid Surfaces for Geometry and Texture Manipulation. IEEE Transactions on Visualization and Computer Graphics, 13(5), 902–913. doi: 10.1109/tvcg.2007.1041

Buehler, P., Everingham, M., Huttenlocher, D.P., Zisserman, A. (2011). Upper body detection and tracking in extended signing sequences. *International journal of computer vision* 95(2), 180

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., Sheikh, Y. (2018). OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields.

Cao, Z., Simon, T., Wei, S.-E., & Sheikh, Y. (2017). Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). doi: 10.1109/cvpr.2017.143

Chai, X., Li, G., Lin, Y., Xu, Z., Tang, Y., Chen, X., Zhou, M. (2013). Sign language recognition and translation with kinect. In: Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition

Creative. 2013. Creative Senz3D™ Interactive Gesture Camera Available for Online Order from 5 September. Accessed on February 18, 2020. https://sg.creative.com/corporate/pressroom?id=13377

Dreuw, P., Rybach, D., Deselaers, T., Zahedi, M., Ney, H. (2007). Speech recognition techniques for a sign language recognition system. In: *Proceedings of the 8th Annual Conference of the International Speech Communication Association (INTERSPEECH),* Antwerp, Belgium.

Dreuw, P., Neidle, C., Athitsos, V., Sclaroff, S., Ney, H. (2008) Benchmark Databases for Video-Based Automatic Sign Language Recognition. In: *The Sixth International Conference on Language Resources and Evaluation (LREC).* Morocco. May 2008.

Dreuw, P., Stein, D., Deselaers, T., Rybach, D., Zahedi, M., Bungeroth, J., Ney, H. (2008). Spoken language processing techniques for sign language recognition and translation. *Technology and Disability* 20 121–133

Fanelli, G., Gall, J., Romsdorfer, H., Weise, T., & Gool, L. V. (2010). A 3-D Audio-Visual Corpus of Affective Communication. IEEE Transactions on Multimedia, 12(6), 591–598. doi: 10.1109/tmm.2010.2052239

Firman, M. (2016). RGBD Datasets: Past, Present and Future. 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). doi: 10.1109/cvprw.2016.88

Gkigkelos, N. and Goumopoulos, C. (2017). Greek Sign Language vocabulary recognition using Kinect. In *Proceedings of the 21st Pan-Hellenic Conference on Informatics (PCI 2017).* Association for Computing Machinery, New York, NY, USA, Article 51, 1–6. DOI: https://doi.org/10.1145/3139367.3139386

Huenerfauth, M., Gale, E., Penly, B., Pillutla, S., Willard, M., Hariharan, D. (2017). Evaluation of Language Feedback Methods for Student Videos of American Sign Language. *ACM Transactions on Accessible Computing* 10, 1, Article 2 (April 2017), 30 pages. DOI: https://doi.org/10.1145/3046788

Huenerfauth, M., Gale, E., Penly, B., Willard, M., Hariharan, D. (2015). Comparing Methods of Displaying Language Feedback for Student Videos of American Sign Language. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility (ASSETS '15).* Association for Computing Machinery, New York, NY, USA, 139–146. DOI: https://doi.org/10.1145/2700648.2809859

Huenerfauth, M. (2020). An Isolated-Signing RGBD Dataset of 100 American Sign Language Signs Produced by Fluent ASL Signers. *Databrary*. Retrieved April 1, 2020 from http://nyu.databrary.org. DOI: http://doi.org/10.17910/b7.1062

Ioannidou, A., Chatzilari, E., Nikolopoulos, S., & Kompatsiaris, I. (2017). Deep Learning Advances in Computer Vision with 3D Data. ACM Computing Surveys, 50(2), 1–38. doi: 10.1145/3042064

Intel. (2020). Intel® RealSense™ Technology. Accessed February 11, 2020. https://www.intel.com/content/www/us/en/architecture-and-technology/realsense-overview.html.

Jiang, Y., Tao, J., Weiquan, Y., Wang, W., Ye, Z. (2014). An isolated sign language recognition system using RGB-D sensor with sparse coding. In: *Proceedings of IEEE 17th International Conference on Computational Science and Engineering.*

Jing, L, Vahdani, E., Huenerfauth, M., Tian, Y. (2019). Recognizing American Sign Language Manual Signs from RGB-D Videos. ArXiv Print: arXiv:1906.02851

Keskin, C., Kra, F., Kara, Y., Akarun, L. (2012). Hand pose estimation and hand shape classification using multi-layered randomized decision forests. In: *In Proceedings of the European Conference on Computer Vision*, pp. 852–863

Koller, O., Forster, J., Ney, H. (2015). Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding.* 141. 108-125. 10.1016/j.cviu.2015.09.013.

Kumar, P., Gauba, H., Roy, P., Dogra, D. (2017). Coupled HMM-based multi-sensor data fusion for sign language recognition. *Pattern Recognition Letters*. 86. 1-8. 10.1016/j.patrec.2016.12.004.

Martnez, A.M., Wilbur, R.B., Shay, R., Kak, A.C. (2002). The RVL-SLLL ASL database. In: *Proceedings of IEEE International Conference Multimodal Interfaces.*

Mehrotra, K., Godbole, A., Belhe, S. (2015). Indian Sign Language recognition using Kinect sensor. In: *Proceedings of the International Conference Image Analysis and Recognition*, pp. 528–535

Microsoft. (2014). Kinect for Windows SDK. Accessed on February 18, 2020. https://docs.microsoft.com/en-us/previous-versions/windows/kinect/dn799271(v=ieb.10)

Microsoft. (2020). Azure Kinect DK – Develop AI Models: Microsoft Azure, Accessed February 11, 2020. http://azure.microsoft.com/en-us/services/kinect-dk/.

Min, R., Kose, N., & Dugelay, J.-L. (2014). KinectFaceDB: A Kinect Database for Face Recognition. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 44(11), 1534–1548. doi: 10.1109/tsmc.2014.2331215

Neidle, C. (2002) SignStream™: A Database Tool for Research on Visual-Gestural Language. In Brita Bergman, Penny Boyes-Braem, Thomas Hanke, and Elena Pizzuto, eds., Sign Transcription and Database Storage of Sign Information, a special issue of Sign Language and Linguistics 4 (2001):1/2, pp. 203-214.

Neidle, C., S. Sclaroff, and V. Athitsos (2001) SignStream™: A Tool for Linguistic and Computer Vision Research on Visual-Gestural Language Data. *Behavior Research Methods, Instruments, and Computers* 33:3, pp. 311-320.

Neidle, C., Vogler, C. (2012). A new web interface to facilitate access to corpora: Development of the ASLLRP data access interface (DAI). In: *Proceedings of the 5th Workshop on the Representation and Processing*

of Sign Languages: Interactions between Corpus and Lexicon, LREC (2012)

Pugeault, N., Bowden, R. (2011). Spelling it out: Real-time ASL fingerspelling recognition. In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 1114–1119

Ren, Z., Yuan, J., Meng, J., Zhang, Z. (2013). Robust part-based hand gesture recognition using Kinect sensor. *IEEE Transactions on Multimedia* 15, 1110–1120

Shah, U., Seita, M, and Huenerfauth, M. (2019). Evaluation of User-Interface Designs for Educational Feedback Software for ASL Students. In: *Antona M., Stephanidis C. (eds) Universal Access in Human-Computer Interaction. Theory, Methods and Tools. HCII 2019. Lecture Notes in Computer Science*, vol 11572. Springer, Cham. DOI: https://doi.org/10.1007/978-3-030-23560-4_37

Simon, T., Joo, H., Matthews, I.A., & Sheikh, Y. (2017). Hand Keypoint Detection in Single Images Using Multiview Bootstrapping. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 4645-4653.

USTC. (2019). University of Science and SLR Group Technology of China, Multimedia Computing & Communication. Chinese Sign Language Recognition Dataset. Accessed February 3, 2020. http://home.ustc.edu.cn/~pjh/openresources/cslr-dataset-2015/index.html

Wei, S., Ramakrishna, V., Kanade, T., & Sheikh, Y. (2016). Convolutional Pose Machines. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4724-4732.

Xie, B., He, X., Li, Y. (2018). RGB-D static gesture recognition based on convolutional neural network. *The Journal of Engineering*, vol. 2018, no. 16, pp. 1515-1520, 11 2018. DOI: 10.1049/joe.2018.8327

Yang, H.D. (2015). Sign language recognition with the Kinect sensor based on conditional random fields. *Sensors* 15, 135–147

Ye, Y., Tian, Y., Huenerfauth, M., Liu., J. (2018). Recognizing American Sign Language Gestures from within Continuous Videos. In *Proceedings of the 8th IEEE International Workshop on Analysis and Modeling of Faces and Gestures (AMFG), The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops,* 2018, pp. 2064-2073.

Zafrulla, Z., Brashear, H., Starner, T., Hamilton H., Presti, P. (2011). American Sign Language recognition with the Kinect. In: In *Proceedings of the International Conference on Multimodal Interfaces*, pp. 279–286

Zahedi, M., Dreuw, P., Rybach, D., Deselaers, T., Bungeroth, J., Ney, H. (2006). Continuous sign language recognition - approaches from speech recognition and available data resources. In: *LREC Workshop on the Representation and Processing of Sign Languages: Lexicographic Matters and Didactic Scenarios,* Genoa, Italy, pp. 21–24

Zhang, C., Tian, Y., Huenerfauth, M. (2016). Multi-Modality American Sign Language Recognition. In *Proceedings of the IEEE International Conference on Image Processing (ICIP 2016)*, Phoenix, Arizona, USA. DOI: https://doi.org/10.1109/ICIP.2016.7532886

8. Language Resource References

Huenerfauth, M. (2020). An Isolated-Signing RGBD Dataset of 100 American Sign Language Signs Produced by Fluent ASL Signers. *Databrary*. Retrieved April 1, 2020 from http://nyu.databrary.org. DOI: http://doi.org/10.17910/b7.1062

Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 95–100
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

Approaches to the Anonymisation of Sign Language Corpora

Amy Isard
Department of Languages, Literature and Media
University of Hamburg
amy.isard@uni-hamburg.de

Abstract

In this paper we survey the state of the art for the anonymisation of sign language corpora. We begin by exploring the motivations behind anonymisation and the close connection with the issue of ethics and informed consent for corpus participants. We detail how the the names which should be anonymised can be identified. We then describe the processes which can be used to anonymise both the video and the annotations belonging to a corpus, and the variety of ways in which these can be carried out. We provide examples for all of these processes from three sign language corpora in which anonymisation of the data has been performed.
Keywords: sign language corpora, corpus anonymisation

1. Introduction

The purpose of anonymisation is to ensure that no personal information is shared for which the person concerned has not given their informed consent. The discussion of what exactly informed consent is, and how to obtain it, is not a simple one (Crasborn, 2010; Rock, 2001; McEnery and Hardie, 2011; Singleton et al., 2014; Schembri et al., 2013). The issues vary depending on among other things the size of the community in which the corpus is collected, the nature of the corpus content and the technological background of the subjects, and it is important to consult the subjects about what they would find appropriate. When describing data collection with the shared signing community in Adamorobe, Kusters (2012, p. 32) observes: "As for anonymity it appeared that people were happy for me to use their real names. The idea of changing their names in 'a book that is about them', seemed very odd to them." Singleton et al. (2014, supplementary material) asked Deaf focus group participants for suggestions about how to use material in research presentations while maintaining anonymity, and they suggested the use of avatars or actors to reproduce the data, or digital editing which could obscure the subject's identity.

Conversations in sign language corpora also often contain mentions of third parties, who are known to the corpus participants but have not been asked for or given any kind of consent for information about them to be shared publicly. Particularly when small communities are involved, it is often easy to identify a person from minimal amounts of information, and care should therefore be taken to obscure as much of this information as possible if videos and annotations are going to be available to the public. Before any analysis or annotation work is carried out on a corpus, participants should always be given a copy of their own recordings and allowed the further opportunity to refuse consent for all or any parts of the recordings to be shown or used in any way.

The process of anonymisation is expensive and time-consuming, and many corpus projects have taken the decision to publicly release only parts of the data where no personal information is revealed, or to ensure that informed consent has been acquired to the best standard possible, and/or that anyone who has access to the data has signed a confidentiality agreement and understands exactly how the data may be used for further research.

In this paper, we describe what the options are once the decision to carry out anonymisation has been taken, and various ways in which these can be implemented. Throughout the rest of the paper, examples of the anonymisation processes and techniques used by the three corpora briefly described below will be used.

The DGS Corpus is a corpus of German Sign Language (DGS). It consists of 560 hours of video dialogues, and about 50 hours has been made available as the Public DGS Corpus[1] (Jahn et al., 2018). The data was elicited using 18 different tasks, some of which involved free conversation where personal information about third parties was sometimes mentioned. The Public DGS Corpus video and annotations have been anonymised to remove references to which would allow the identification of third parties.

The NGT Corpus is a corpus of Dutch Sign Language (NGT). It consists of dialogues between 92 participants and is available online[2] (Crasborn and Zwitserlood, 2008). A number of different elicitation tasks were used and some of the conversations involve references which could identify third parties. The available annotations have been anonymised but the video has not.

The Rudge Corpus is a small corpus of British Sign Language (BSL) collected by Luke Rudge for his PhD thesis on the topic of the use of Systemic Functional Grammar in the analysis of BSL (Rudge, 2018). There were 12 participants who gave pre-prepared presentations about a prominent period in their lives, which sometimes revealed personal information. The videos and annotations have been anonymised but they are not publicly available.

2. What to Anonymise

In sign language corpora, it is impossible to completely anonymise the video data, because both the face and hands

[1] http://ling.meine-dgs.de
[2] https://www.ru.nl/corpusngten/

of the participants must be fully visible for the content to be understandable (Quer and Steinbach, 2019; Hanke, 2016; Crasborn, 2010). Chen Pichler et al. (2016, page 32) note that: "there appears to be virtually unanimous agreement that total anonymization, long taken as a standard practice for medical data, is not feasible for language data that include audio and/or video components".
Although it is not possible to completely conceal the identities of the participants in a sign language corpus, it is nonetheless necessary to ensure that as few of their personal details are revealed as possible. In addition, care must be taken to obscure personal information of third parties who are mentioned during the dialogue, if it could lead to their identification. These third parties will not have had the opportunity to give their informed consent for any sort of appearance in the corpus.
There are two main situations in which the anonymisation of sign language corpora is carried out:

- anonymisation of a whole corpus for wider distribution to a larger team or outside researchers
- anonymisation of single words or phrases for use in settings such as a conference talk, seminar or sign language dictionary

In both cases, it is first necessary to identify which information needs to be anonymised. In a small corpus it may be possible to make the selection by watching all the videos, but in a larger corpus it maybe helpful to use some automatic processing. The anonymisation of videos is described in Section 3, and of annotations in Section 4.

3. Anonymisation of Video

There are a number of different ways in which video can be anonymised. These can be divided into two categories, those which *conceal* all or part of a video, and those which *reproduce* a video. Concealing can be effected on part or all of a video frame with the use of blurring or pixellation, or by obscuring the image entirely. Reproduction can be carried out by using either an actor or a computer-generated avatar.
These two approaches are generally used for different purposes. Reproduction can conceal the identity of the signers themselves, while concealing preserves the anonymity of third parties by hiding references to people or places. No detailed studies have been published about the extent to which reproduction affects the viewer's understanding of a sign language video, or what level of blurring is necessary to ensure that the movements cannot be distinguished. In the related area of spoken dialogue research, the CASE corpus of Skype dialogues experimented with video anonymisation using *Adobe Premiere* pixel, art, and transformation filters, and chose a contour filter. In control tests, they discovered that when this filter was used, subjects did not recognise themselves (Diemer et al., 2016).

3.1. Concealment

Concealment can be used on just part of the image of a video, and usually over a small time frame. The viewer's experience is not hugely disrupted, as only a sign or two will be concealed. Inevitably some information will be lost, but this can be kept to a minimum. The concealment can be carried out by blackening all or part of the image ()adding one or more black rectangles), or by blurring or pixellating all or part of the image to such an extent that the signing or mouthing is no longer recognisable.

Figure 1: Screenshot from the DGS Corpus, anonymised through blackening with one black rectangle over the mouth and cheeks and another over the right hand and arm and the top right portion of the torso.

3.1.1. Blackening

In the Public DGS Corpus mentions of sensitive information in videos are anonymised by blackening sections of the image (Bleicken et al., 2016). The timings from the annotation tiers (see Section 4) are used to identify the relevant timespan. Experiments were carried out which showed that if the whole timespan was blackened, this invalidated a whole sentence for linguistic analysis, because it disturbed suprasegmental signals. They therefore imposed one or more black rectangles on the image, to cover the mouth, one or both hands and/or the trunk, depending on the position of the sign. Experiments also showed that blackening was less disturbing to viewers than pixellation. OpenPose analysis (Cao et al., 2017) had already been carried out on the corpus (Schulder, 2019), providing machine-readable information on the location of various body parts, such as hands, shoulders, and mouth, so this was used to find the location of the relevant body parts, and the size and shape of the rectangles were then adjusted by hand. An example screenshot is shown in 3.1.1, where the mouth, cheeks, right hand and right arm of the signer have been hidden, along with a portion of the torso in front of which the sign was being performed.

3.1.2. Pixelation

For the Rudge corpus, the author went through the video recordings and noted where participants had signed a proper name of a person, specific location or any other information which could identify a third party. The video was then loaded into editing software such as *Final Cut Pro* or *Adobe After Effects*, and a local blur or pixellation filter was applied to the signer's hands and mouth for the duration of the relevant sign, which was normally only a few tenths of a second during fluent signing (Rudge, personal communication, January 2020). This ensured that any third party information had been removed before the recordings were

passed to other researchers for annotation. No screenshots are available as participants did not give consent for any images to be shown to people outside the initial small research group.

3.2. Reproduction

Reproduction of a corpus can in theory be carried out by either humans or computer-generated avatars. Some corpus examples where human actors have been used are described in Section 3.2.1 and the steps which would be necessary for avatar reproduction in Section 3.2.2.

3.2.1. Actors

For total anonymity, short examples from a corpus can be reproduced by a human actor. In this case complete anonymity is assured, but there are several disadvantages as a result. The process is very labour-intensive, requiring not only the time of the signer but also of a studio and technicians to carry out the recording. In addition, no matter how well the second signer copies the original, some information will be lost. Performativity is a vital part of sign language and it is impossible to fully separate the affective and grammatical functions of facial expressions.

The participants in the Rudge corpus had agreed only to their recordings being seen by the author and a limited number of other researchers who worked on verification of the data. Because the thesis is publicly available, examples used in it were reproduced by the author or another signer, to preserve the anonymity of the original participants (Rudge, 2018 and personal communication, January 2020).

The DGS Corpus is being used in the compilation of a Dictionary of German Sign Language and the preference is to use examples taken directly from the corpus, for the reasons discussed in detail in Langer et al. (2018). However, in very occasional cases where the dictionary compilers want to use an example which contains personal information about a third party, they re-record the example with a signing model and replace any personal names in the re-recording and the associated translation with a common German family name.

3.2.2. Avatars

In practice, although avatars have been improving rapidly in quality, no large-scale avatar reproduction has been carried out. A survey of the state of the art in sign language avatars can be found in (Bragg et al., 2019). The are a number of technical problems with the use of avatars for sign language, and some of these are related to the process of creating the content and ensuring that the correct manual and non-manual gestures are created. In the case of reproduction these particular issues are avoided, because the data for the avatar comes directly from the original videos. The problems of designing avatars which are acceptable to the Deaf community in terms of appearance and comprehensibility remain, and it is essential that the acceptability of avatars be systematically reviewed and assessed before they are used (Kipp et al., 2011).

In order to use avatars for reproduction, the original videos must first be processed using pose estimation software, which can identify particular body parts including hands, arms, and facial features. A visual representation of an OpenPose analysis from the DGS corpus (Schulder, 2019) is shown in Figure 3.2.2, illustrating the keypoints identified by the software and lines between the points to indicate logical connections between them. However, OpenPose only produces two-dimensional images, and additional (extremely time and resource intensive) processing is required to reconstruct three-dimensional images (Xiang et al., 2019). The resulting machine-readable information on the location of various body parts could then be used to animate an avatar which would reproduce the desired data, but as far as we are aware, no sign language avatar has so far been tested on this output.

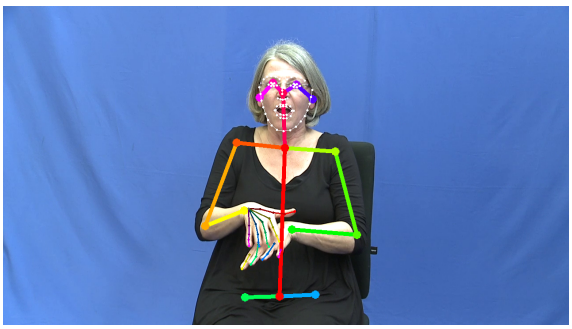


Figure 2: Visual representation of the pose information provided by OpenPose, computed for a video from the DGS-Korpus project. Sets of keypoints are generated for the body, the face and each hand. Lines between the points are added to the visual representation to indicate the logical connection between individual keypoints.

3.2.3. OpenPose data

If OpenPose data are made publicly available, they must also be anonymised, to the same level as the videos on which they were based. If the data were later used, for example to animate an avatar, they could make personal information visible. OpenPose data are available to download as part of the Public DGS Corpus (Schulder, 2019), and they have been anonymised to remove keypoints for timespans which were previously chosen for anonymisation as described in Section 4.1. It is possible to differentiate between keypoints which have been anonymised and those which are missing because the body part is temporarily hidden (for example when a person puts a hand behind their head), so that if the OpenPose data were used to animate an avatar, anonymised keypoints could be covered by a black square, as with video blackening (see Section 3.1.1).

4. Anonymisation of Annotations

Before the anonymisation of annotations can be carried out, the sensitive names must first be identified. In a small corpus, this may have been done by watching the video data, but where many hours of video have been translated and annotated by a team of researchers, automatic methods can also be used.

4.1. Name Identification Methods

For the Rudge corpus, names were found by manual inspection of the videos (see Section 3).
In the NGT corpus, information which had been manually annotated in the gloss and mouth tiers was used to identify names which needed to be anonymised (Crasborn and Bank, 2015).
The DGS-Korpus project tested a subset of the DGS corpus to see how reliable different techniques were for finding sensitive items which should be anonymised (Bleicken et al., 2016). Because German translations had already been carried out, they could use computational linguistic tools for German which are available through Weblicht (Hinrichs et al., 2010) as pre-defined chains.
They used four approaches and compared the results for each to a ground truth defined as the sum of the names correctly identified by each technique. The four approaches which they used were:

- Manual inspection of the videos by a deaf annotator who was asked to mark every occurrence of a name
- Extraction of potential names from the annotations, which were then checked against the German translations; when a match was found, a manual inspection was carried out
- Use of named entity recognition on the German translations
- Checking mouthing annotations and translations against name lists

When comparing the final outcomes of all the methods, they found that the most effective process was to combine the automatic methods with a one-pass manual inspection.
The DGS-Korpus project found that they were more conservative in their selection of data which needed to be anonymised than the participants themselves had been after reviewing their own recordings. They decided therefore that it was unfair to make the participants entirely responsible for these decisions, and better to be more cautious, and carry out more anonymisation rather than less, in an effort to prevent any identifiable information on third parties being released accidentally.
Once the names to be anonymised have been identified, the actual anonymisation can be done using either *categorisation* or *pseudonymisation*. Pseudonymisation involves the use of replacement names (Section 4.2). In categorisation, a name is usually replaced by a string indicating the type of proper name plus a numeric identifier, so that subsequent mentions in the same dialogue can be seen to be referring to the same entity (Section 4.3).

4.2. Pseudonymisation

When pseudonymisation is carried out, the pseudonyms can be chosen to match the original names on as many levels as desired. This could for example involve choosing replacement cities of approximately the same size, or family names which originate from the same geographical region. Anonymisation with pseudonyms was for example carried out in the spoken German FOLK corpus (Schmidt, 2016; Winterscheid, 2015). One disadvantage of this approach is that it can be very time consuming as time must be spent choosing replacement names and making sure that they fit all of the chosen criteria. There are currently no sign language corpora for which a description of anonymisation using pseudonyms is available. Issues to consider would include the question of how to define "similar" names in terms of sign language phonology.

4.3. Categorisation

Categorisation is a quicker and simpler process than pseudonymisation because it is only necessary to identify the type of a proper name in order to create its replacement.
In the NGT corpus, glosses and annotation tiers are anonymised so that it will not be possible for anyone to make a simple automatic search for names. All glosses which refer to participants and other people who are not considered to be in the public domain are replaced by the type `*NAMESIGN`. In mouthing and translation tiers, they are replaced by the type `*eigennaam` (Crasborn and Bank, 2015).
In the Rudge corpus, the timestamps from the manual analysis of the video data (see Section 3.1.2) were used to find places in the translation and annotations where names, locations and other personal data needed to be anonymised. They were replaced with types such as `[NAME]` or `[LOCATION]`. If there were multiple instances of anonymisation in the same clause or in quick succession, a suffix was added of the form `[NAME-a]`, `[NAME-b]`, etc. so that any following indicating verbs or signs requiring more complex spatio-kinetic features (e.g., placement in the signing space) could still be understood in spite of the visual noise (Rudge, 2018 and personal communication, January 2020).
The DGS-Korpus project examined each person name to determine whether it belonged to someone for whom information is already available in the public domain, such as television personalities or politicians, whose names would not then be anonymised. They also defined a population threshold above which places were considered to be large enough to not require anonymisation. Proper names in the translation and mouthing annotations, and most of the gloss tier, were replaced by numbered placeholders of the form `Name#1`, `Name#2`, etc. so that it is still possible to tell when the same person or place is referred to more than once.

5. Final Thoughts

It must always be kept in mind that in a large corpus it is basically impossible to ensure that all possible identifiable information has been removed, and that this must be made clear to the participants as part of the process of obtaining informed consent. For example, in one dialogue from the Public DGS corpus (English translation shown below), a place name is anonymised, but two sentences later it is mentioned that it is the previous residence of a princess from the 18th century who, as a person in the public eye, would not normally have her name anonymised:

My hometown Place#1 also has a small tourist attraction.
There used to be a castle right where the German Catholic Church is located today.
The Austrian princess Elisabeth used to live there.

It would therefore be theoretically possible for someone who comes from the same area or has a thorough knowledge of the history of the region to figure out the name of the participant's home town. To avoid this, the name of the princess would then also have to be anonymised, and possibly even her nationality, but at some point a decision has to be made about how far to continue the process, and in this case, it was decided that the name of the princess would not be anonymised.

6. Acknowledgements

This work was supported by the BMBF (German Federal Ministry of Education and Research) Project QUEST: Quality-Established[3].

7. Bibliographical References

Bleicken, J., Hanke, T., Salden, U., and Wagner, S. (2016). Using a Language Technology Infrastructure for German in order to Anonymize German Sign Language Corpus Data. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3303–3306, Portorož, Slovenia.

Bragg, D., Koller, O., Bellard, M., Berke, L., Boudreault, P., Braffort, A., Caselli, N., Huenerfauth, M., Kacorri, H., Verhoef, T., Vogler, C., and Ringel Morris, M. (2019). Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, pages 16–31, Pittsburgh, PA, USA.

Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. (2017). Realtime multi-person 2D pose estimation using part affinity fields. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7291–7299, Honolulu, HI, USA.

Chen Pichler, Deborah, D., Hochgesang, J., Simons, Doreen, D., and Lillo-Martin, Diane, D. (2016). Community Input on Re-consenting for Data Sharing. In *Proceedings of the Seventh Workshop on the Representation and Processing of Sign Languages: Corpus Processing at LREC 2016*, Portorož, Slovenia.

Crasborn, O. and Bank, R. (2015). Corpus NGT Anonymisation Protocol. `https://www.academia.edu/40438732/Corpus_NGT_Anonymisation_Protocol`.

Crasborn, O. A. and Zwitserlood, I. E. P. (2008). The Corpus NGT: An Online Corpus for Professionals and Laymen. In *Proceedings of the Third Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora at LREC 2008*, pages 44–49, Marrakech, Morocco.

[3] `https://www.slm.uni-hamburg.de/en/ifuu/forschung/forschungsprojekte/quest.html`

Crasborn, O. (2010). What Does "Informed Consent" Mean in the Internet Age? Publishing Sign Language Corpora as Open Content. *Sign Language Studies*, 10(2):276–290.

Diemer, S., Brunner, M.-L., and Schmidt, S. (2016). Compiling computer-mediated spoken language corpora: Key issues and recommendations. *International Journal of Corpus Linguistics*, 21(3):348–371.

Hanke, T. (2016). Towards a Visual Sign Language Corpus Linguistics. In *Proceedings of the Seventh Workshop on the Representation and Processing of Sign Languages: Corpus Mining at LREC 2016*, pages 89–92, Portorož, Slovenia.

Hinrichs, E., Hinrichs, M., and Zastrow, T. (2010). WebLicht: Web-Based LRT Services for German. In *Proceedings of the ACL 2010 System Demonstrations*, pages 25–29, Uppsala, Sweden.

Jahn, E., Konrad, R., Langer, G., Wagner, S., and Hanke, T. (2018). Publishing DGS corpus data: Different Formats for Different Needs. In *Proceedings of the Eighth Workshop on the Representation and Processing of Sign Languages: Involving the Language Community at LREC 2018*, pages 83–90, Miyazaki, Japan.

Kipp, M., Heloir, A., and Nguyen, Q. (2011). Sign Language Avatars: Animation and Comprehensibility. In Hannes Högni Vilhjálmsson, et al., editors, *Intelligent Virtual Agents*, Lecture Notes in Computer Science, pages 113–126, Berlin, Heidelberg. Springer.

Kusters, A. (2012). Being a deaf white anthropologist in Adamorobe: Some ethical and methodological issues. In *Sign Languages in Village Communities: Anthropological and Linguistic Insights*, pages 27–52. De Gruyter Mouton, Berlin, Boston.

Langer, G., Müller, A., Wähl, S., and Bleicken, J. (2018). Authentic Examples in a Corpus-Based Sign Language Dictionary – Why and How. In *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts.*, pages 483–497, Ljubljana, Slovenia.

McEnery, T. and Hardie, A. (2011). *Corpus Linguistics: Method, Theory and Practice*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Quer, J. and Steinbach, M. (2019). Handling Sign Language Data: The Impact of Modality. *Frontiers in Psychology*, 10.

Rock, F. (2001). Policy and Practice in the Anonymisation of Linguistic Data. *International Journal of Corpus Linguistics*, 6(1):1–26.

Rudge, L. A. (2018). *Analysing British Sign Language through the Lens of Systemic Functional Linguistics*. Ph.D. thesis, University of the West of England. `https://uwe-repository.worktribe.com/output/863200`.

Schembri, A., Fenlon, J., Rentelis, R., Reynolds, S., and Cormier, K. (2013). Building the British Sign Language Corpus. *Language Documentation & Conservation*, 7:136–154.

Schmidt, T. (2016). Construction and Dissemination of a Corpus of Spoken Interaction - Tools and Workflows

in the FOLK project. *Journal for Language Technology and Computational Linguistics*, 31(1):127–154.

Schulder, M. (2019). OpenPose in the Public DGS Corpus. Project Note AP06-2019-01, Institute for German Sign Language, Hamburg University, Hamburg, Germany. https://www.sign-lang.uni-hamburg.de/dgs-korpus/arbeitspapiere/AP06-2019-01.html.

Singleton, J. L., Jones, G., and Hanumantha, S. (2014). Toward Ethical Research Practice With Deaf Participants. *Journal of Empirical Research on Human Research Ethics*, 9(3):59–66.

Winterscheid, J. (2015). Maskierung. Working Paper, Institut für Deutsche Sprache, Mannheim. https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/3904/file/Winterscheid_Maskierung_2015.pdf.

Xiang, D., Joo, H., and Sheikh, Y. (2019). Monocular total capture: Posing face, body, and hands in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10965–10974, Long Beach, CA, USA.

Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 101–106
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

Sign Language Motion Capture Dataset for Data-driven Synthesis

Pavel Jedlička, Zdeněk Krňoul, Jakub Kanis, Miloš Železný
NTIS - New Technologies for the Information Society,
Faculty of Applied Sciences, University of West Bohemia
Univerzitní 8, 306 14 Pilsen, Czech Republic.
{jedlicka, zdkrnoul, jkanis, zelezny}@ntis.zcu.cz

Abstract

This paper presents a new 3D motion capture dataset of Czech Sign Language (CSE). Its main purpose is to provide the data for further analysis and data-based automatic synthesis of CSE utterances. The content of the data in the given limited domain of weather forecasts was carefully selected by the CSE linguists to provide the necessary utterances needed to produce any new weather forecast. The dataset was recorded using the state-of-the-art motion capture (MoCap) technology to provide the most precise trajectories of the motion. In general, MoCap is a device capable of accurate recording of motion directly in 3D space. The data contains trajectories of body, arms, hands and face markers recorded at once to provide consistent data without the need for the time alignment.

Keywords: Sign Language, Motion Capture, Dataset

1. Introduction

Sign language (SL) is a way of communication that utilizes the movement of a human body. It uses manual, facial, and other body movements to express information. SL is a basic communication system of deaf people and it is often their natural way of communication. According to (Naert et al., 2017), deaf people are often facing problem using written language (based on the spoken language), because it uses the different grammatical rules, and the nature and the spatial organization of linguistic concepts as well. However, most information in the media or the Internet is available in the spoken or the written form. Thus it leads to difficulties for deaf people to access the information.

Computer animation techniques have experienced great improvement recently. There have been developed devices dedicated to the recording of a movement in high precision in 3D space. Animations computed from the data recorded in this way are of high quality and accurate, and their usage is increasingly common outside the film and the computer game industry. An artificial avatar is one possible output of such animation. In public television as an example, they use translation made by a signer which is shown in a window added into the screen. However, the avatar technology is more flexible compared to the real SL signer. It has editable content that can be produced more easily than video (no recording studio with camera) and which also preserves the anonymity of the signer. Using an animated artificial avatar with automatic SL synthesis seems to be a good way to improve the actual way of using CSE on TV.

Recently, some approaches based on key-frame techniques and procedural synthesis have been developed. These approaches provide fine control over the movements of the avatar. These avatars are however poorly accepted by the deaf community because of their lack of human-like motion. There are some works that aim to deal with this problem. In (McDonald et al., 2016) for example, authors added noise measured from MoCap data to the rule-based synthesis to improve the performance of the avatar. Data-driven synthesis, on the other hand, preserves the motion of an original SL signer.

In this paper, we introduce, by our best knowledge, the first MoCap dataset of CSE. This dataset consists of both dictionary items and continuous signing. Manual and non-manual components were recorded simultaneously and the setup includes a high number of markers placed on the face, the body and fingers in order to provide precise and synchronous data. As the main purpose of creating this dataset is to develop an automatic SL synthesis, we also suggest the methods for evaluating the synthesized data.

2. Related Work

Most SL datasets are recorded by an optical camera as they are the most affordable device for this purpose and the recording setup is fast. The difference in data output from the MoCap system and video output is that the MoCap system provides 3D data directly and therefore can be more precise. Although, there are techniques developed for the pose estimation from the image or video, e.g. OpenPose (Cao et al., 2017), the 3D precision is in principle lower than the actual 3D pose measuring provided by the MoCap system.

Some datasets using different motion capture techniques were created in recent years. (Lu and Huenerfauth, 2010) recorded American SL using magnetic-based motion capture for hand and finger tracking. The evolution of motion capture datasets collected in French SL is described in (Gibet, 2018). They recorded three MoCap datasets in the last 15 years. All of them contain manual and non-manual components of SL. The project HuGEx (2005) used Cybergloves for recording finger movements and the Vicon MoCap system for the body and the facial movements. The total recording time was 50 minutes. The next project, SignCom (2011) uses the Vicon MoCap system to record all components and the recording time was 60 minutes, but only 6 markers per hand were used for the hand and finger recording. The most recent project Sign3D (2014) has

all components recorded with the Vicon system and the eye gaze was recorded with a head-mounted oculometer (MocapLab MLab 50-W). It has 10 minutes of recorded data.
There is a continual need for a large amount of data to utilize machine learning techniques. Although the quality and size of datasets are increasing, there is still a lack of such data. The usual size of those datasets is between 10 and 60 minutes of recording time.

3. Dataset Design

Our aim is to record the SL dataset usable for automatic synthesis and evaluation of new utterances. In order to synthesize any given utterance, the language domain was limited to the terms used in the weather forecast. The weather forecast domain was also selected because of the availability of reference video recordings of daily forecasts in SL from a recent couple of years. The size of the vocabulary is reasonably limited for our purposes.
There are some differences in SL expressions depending on the location due to different dialects of CSE, therefore, we used the video source provided by the Czech national television because the used signs are considered as well understandable and recognizable to most of the audience.
CSE linguist experts selected 36 weather forecasts broadcasted throughout the year in order to provide different expressions needed for weather forecasts in different seasons to provide all the necessary data for further synthesis of any weather forecast in the future.

4. Recording Setup

The Motion capture (MoCap) recording is the process of recording the movements using specialized devices in order to reconstruct motions in the 3D space during the time. There are different approaches for data acquisition using MoCap techniques and there are also devices dedicated to the MoCap recording of different body parts. We did some experimental recordings using a different variation of devices such as Cybergloves2 for finger and VICON Cara for facial recording (Krňoul et al., 2016). The main problem with the usage of such devices was signer's discomfort and limitations to performed movements (e.g. tight gloves reduce free movement of fingers, Cara devices camera placement denies finger-face interactions). Another issue was synchronization and calibration (data alignment in general) of different devices as described in (Huenerfauth et al., 2008) and (Krňoul et al., 2016).
Recording all modalities (arm, hand pose, and facial movement) using one device emerged as the best solution. In our solution using an optical-based MoCap system, the signer is equipped with lightweight markers only, and there is no need for merging data together. The only limitation is that the optical-based approach needs a clear line of view from cameras to markers and, therefore, is sensitive to occlusions of body parts. A large number of cameras are needed as well as their precise placement, for such a complex movement like SL utterances.

4.1. Motion Capture Setup

We used the optical-based MoCap system consisting of 18 VICON cameras (8xT-20, 4xT-10, 6xVero) for dataset recording and one RGB camera as referential and two Kinects v2 for additional data acquisition. MoCap recording frequency was 120Hz. The placement of cameras shown in Figure 1 was developed to cover the place in front of the signer in order to avoid occlusions as much as possible and in order to focus on facial expressions. Camera placement was also adjusted for the particular signer to reduce gaps in trajectories caused by occlusions.

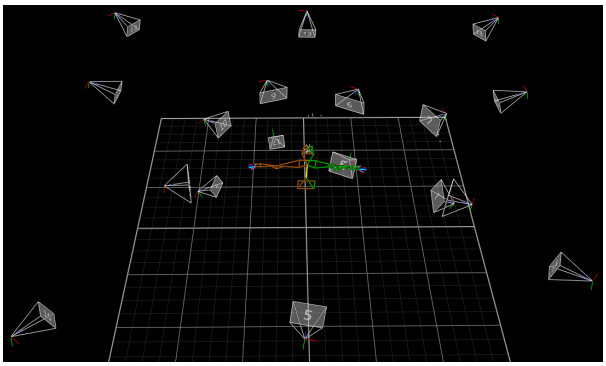


Figure 1: Visualization of MoCap camera layout. View from back and above, the signer is in the middle.

4.2. Subject Setup

The markers placed on the face and fingers were selected to cause minimal disturbance to the signer. We used different marker sizes and shapes for different body parts (see Table 1 and Figure 2). We tracked the upper body and arms by a pair of markers placed on the axis of joints completed by some referential markers. The positions of markers on the face were selected to follow facial muscles and wrinkles. We used 8mm spherical markers around the face, 4 mm hemispherical markers for facial features with the exception of nasolabial folds with 2.5 mm hemispherical markers. The eye gaze and eyelid movement were not tracked by the MoCap device, but it can be obtained from the reference video. Two markers for palm tracking are placed on the index and small finger metacarpals. We tracked fingers using three 4 mm hemispherical markers per finger placed in the middle of each finger phalanx and thumb metacarpals.

	marker diameter [mm]	marker count
Body, arms, hands	8 - 14	33
Fingers	4	30
Face	2.5 - 8	46
Total	2.5 - 14	109

Table 1: Maker sizes and count per segment.

5. Dataset Parameters

We have recorded approximately 30 minutes of continuous signing ($>$ 200000 frames) and 12 minutes of dictionary items. All data were recorded by one expert CSE signer,

Figure 2: Signer marker setup.

who was monitored by another CSE expert during the process. The dataset contains 36 weather forecasts. On average, each such forecast is 30 seconds long and contains 35 glosses. The dictionary contains 318 different glosses. Those dictionary items are single utterances surrounded by the posture with loose hands and arms (a rest pose) in order not to be affected by any context.

Dataset processing is a very demanding work both in terms of time and demands for expert annotation and MoCap data postprocessing. MoCap data have to be processed in order to ensure proper labeling of each marker and to fill eventual gaps in marker trajectories. The next step of MoCap data processing is to solve the marker trajectories (Figure fig:MarkerSetup) to the form of the skeleton model shown in Figure 5. Solving provides data in the angular domain of each body part. Those data can be used directly for the animation.

Another important step in the processing of the dataset is the annotation of content. We used the well-known Elan annotation tool for this purpose, see (Crasborn and Sloetjes, 2008). The reference video of data was used for the annotation as it provides the possibility to annotate the data without need of rendering the MoCap data but it lacks precision because of lower frame-rate (120 fps MoCap vs. 25 fps video). This annotation was made by the CSE native signer. It contains time stamps dividing the data into different signs, transitions between signs and rest pose in one-tier, see Figure 3. The aim of this annotation is to roughly capture those moments of change and it will be used as initialization for a data-driven segmentation/synthesis process. Although annotation is still in progress, almost 80% is already done.

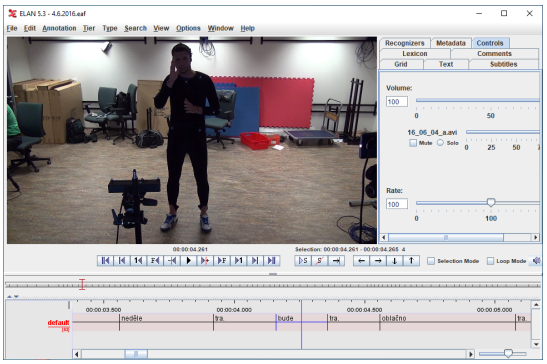


Figure 3: Annotation in ELAN.

6. Data and Synthesis Evaluation

The best way and till now mostly used method for expressing the quality or comparing the similarity of two signs is using subjective evaluating by SL native signers. However, this evaluation is both time and human resources demanding process and moreover usually more than one person is needed for the subjectivity of the evaluation, see (Huenerfauth et al., 2008).

The popularity of automatic and machine learning techniques utilization for data-processing related tasks increased in recent years. An objective criterion in the form of a cost function is crucial for such techniques but it is usually not trivial to choose one. The purpose of such a function is not to replace the human evaluation of the synthesis result, but to provide a proper cost function for machine learning techniques as they need fast evaluation during training process.

The data provided by the MoCap recording are trajectories of all markers. The advantage of such data is direct information of the positions in the 3D space but the human body topology (skeleton) may not be respected in such representation. On the other hand, angular trajectories of bones are bound to the exact human body topology. The topology of a signer is constant during the time. This can improve the consistency of the data if signs from single signer are compared. In both cases, one frame can be considered as a vector of values and the duration of two similar utterances can differ, although the meaning is the same. The signs and utterances are the time-sequences of these vectors.

The usual metrics (among the others) for evaluating difference/similarity between two single vectors $p = (p_0, p_1, ..., p_i, ...p_N)$ and $q = (q_0, q_1, ..., q_i, ...q_N)$ of the same length N are:

- Euclidean distance:

$$d = \sqrt{\sum_{i=0}^{N} (q_i - p_i)^2}, \quad (1)$$

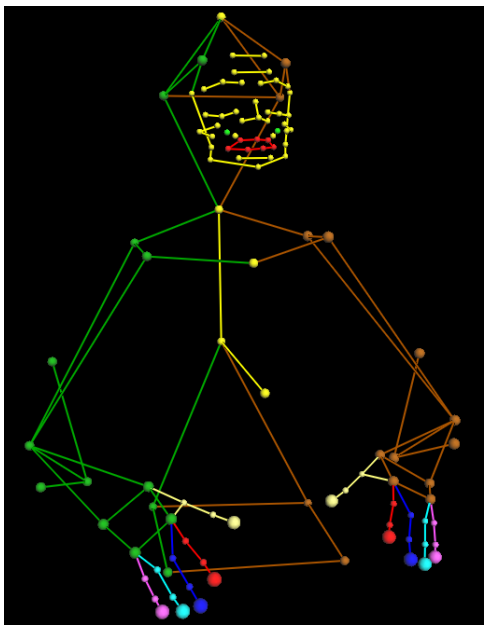


Figure 4: Marker setup (data visualization).

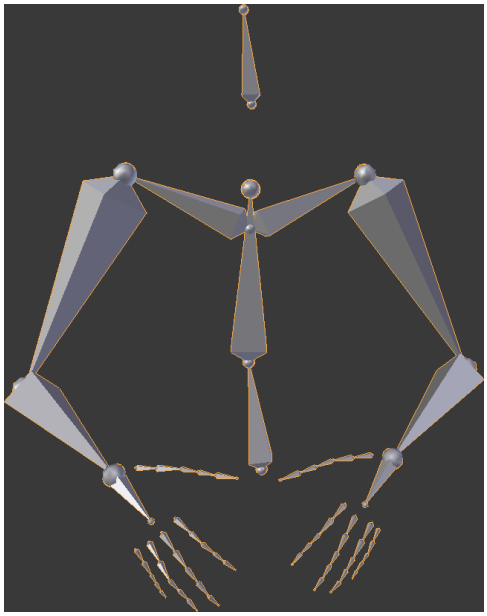


Figure 5: Model visualization.

- Root mean square error (RMSE):

$$d = \sqrt{\frac{\sum_{i=0}^{N}(p_i - q_i)^2}{N}}, \tag{2}$$

- Correlation coefficients (Corr):

$$d = \frac{\sum_{i=0}^{N}(p_i - \overline{p})(q_i - \overline{q})}{\sqrt{\sum_{i=0}^{N}(p_i - \overline{p})^2 \sum_{i=0}^{N}(q_i - \overline{q})^2}}, \tag{3}$$

where $\overline{p}$ and $\overline{q}$ are mean values of p and q respectively.

The time component of the data (the time-sequence of the vectors) can be addressed by the following approaches. One of them is a time alignment in the form of re-sampling the time-sequence of two compared components to the same length and then measure the distance. In (Sedmidubsky et al., 2018) they used normalization for motion data comparison for query purposes in the form of the time axis movement sequence normalization and Euclidean distance for each motion.
Dynamic time warping (Berndt and Clifford, 1994) (DTW) is commonly used algorithm for the time-series comparison. This method computes the best per frame alignment in terms of the chosen distance. It provides us a possibility to get minimal distance of two time-sequence with different lengths, for example two utterances with different signing pace. The computed DTW distance d_DTW is a minimal distance with the optimal time alignment of sequences p and q, $path$ describes the alignment of the vectors:

$$d_{DTW}, path = DTW(p, q). \tag{4}$$

We tested the DTW algorithm with the Euclidean distance (1) for measuring a distance between two different signs and between different instances of the same sign. We limited this test for the signs with meanings "one", "two", "three", "four", and "five", both from the dictionary and the continuous signing and compared measured distances between signs with the same meaning (different instance of the same sign) and different signs (all instances of other signs from the same test-set). The DTW distance was measured between two signs, the distance was normalized to the vector size and the length of the DTW path, so the distance is independent on the skeleton complexity and duration of the sequence. The normalized DTW $d_{normDTW}$ distance is defined as:

$$d_{normDTW} = \frac{d_{DTW}}{M \cdot N}, \tag{5}$$

where M is the length of the $path$ from DTW algorithm and N is the number of channels of the data.

Sign	distances [deg] (same meaning)	distances [deg] (different meaning)
"one"	0.84 - 1.79	2.49 - 8.67
"two"	0.45 - 1.29	2.49 - 7.08
"three"	1.18	2.54 - 5.80
"four"	0.33 - 0.85	3.24 - 8.67
"five"	0.33 - 0.85	2.49 - 7.78

Table 2: Normalized DTW distances between signs (handshapes only).

The Euclidean distances of angular trajectories computed using DTW are summarized in Tables 2 and 3 for handshape only and for the whole body (hand included) respectively. The tested signs (numbers from 1 to 5) were chosen because they are very similar and differs only in the handshape. The signs are compared to other instances with the same meaning and to all instances of all different signs (e.g.

Sign	distances [deg] (same meaning)	distances [deg] (different meaning)
"one"	2.30 - 3.25	3.08 - 6.90
"two"	1.04 - 3.59	2.74 - 6.37
"three"	2.58	2.94 - 5.42
"four"	0.89 - 3.28	2.73 - 6.90
"five"	1.10 - 2.07	3.13 - 5.57

Table 3: Normalized DTW distances between signs (whole body without face).

all instances with the meaning "one" are compared to all other instances with the same meaning and to all instances with different meanings such as "two", "three", ...). According to the results in Table 3, using normalized DTW distance for raw trajectories of the angular representation seems to have the ability to objectively measure the difference between signs, because the distance is generally lower for the signs with the same meaning than others.
In case of the hand-shapes (Table 2, there seems to be the ability to not only measure the distances between signs with the same meanings but also to distinct different signs completely.
We suggest some approaches to improve the evaluation of distances calculated by DTW. We can use different weights for the distance measure for different bones based on its corresponding importance for the signs distinction. We can also use trajectories of different body parts to compare signs components separately. For example, compare handshapes, palm orientation and location with their counterparts respectively to enable more precise modeling of SL grammar such as classifiers, the co-occurrence of manual and non-manual, etc.

7. Experiments

7.1. Methods

We propose the following baseline technique for the SL utterance synthesis. The purpose of this baseline is not to solve the synthesis problem itself but to provide a reference algorithm and performance for further developed and more sophisticated techniques. We assemble the utterance from dictionary item trajectories for each sign. Then we compute trajectories of transition movement between these signs. We set the fixed length for all transitions as the average length of all transitions in our dataset. We interpolated the transition trajectory for each joint by the cubic spline. For evaluation, we compared the synthesized utterance with the utterance captured in the continuous signing by the normalized DTW with Euclidean distance.

7.2. Results

We selected a pair of utterances that have more appearances in the dataset in order to provide a comparison with a reference.

- Utterance 1: "zima-hory-kolem" (literal translation: cold-hills-approximately). Confusion matrix is shown in Table 4
- Utterance 2: "pocasi-zitra-bude" (literal translation: weather-tomorrow-will be). Confusion matrix is shown in Table 5

In confusion matrices (Tables 4 and 5), we can see the normalized DTW distances of the synthesized utterance compared to utterances with the same meaning that appear in continuous signing. For reference, we added a comparison with the utterance with other meaning.

	synth	*appear1*	*appear2*	*appear3*	*other*
synth	0	2.58	2.69	2.82	5.28
appear1	2.58	0	1.03	1.27	6.14
appear2	2.69	1.03	0	1.41	6.19
appear3	2.82	1.27	1.41	0	6.62
other	5.28	6.14	6.19	6.62	0

Table 4: Confusion matrix of normalized DTW distances for utterance 1. Synthesised data (*synth*), compared with real data (*appear1-3*) and *other* utterance with different meaning.

	synth	*appear1*	*appear2*	*appear3*	*other*
synth	0	1.51	1.43	1.61	5.28
appear1	1.51	0	0.62	0.71	4.69
appear2	1.43	0.62	0	0.82	4.84
appear3	1.61	0.71	0.82	0	4.60
other	5.28	4.69	4.84	4.60	0

Table 5: Confusion matrix of normalized DTW distances for utterance 2. Synthesised data (*synth*), compared with real data (*appear1-3*) and *other* utterance with different meaning.

The comparison of the normalized DTW distances shows larger differences between synthesized utterance and examples from continuous data then among the continuous data. We can also distinct different utterances from each other. The difference between synthesized data and examples from continuous data can be caused by various reasons. We try to explain some of those in the following discussion.

8. Discussion

There is a difference in the pacing and the method of signing for signs in the dictionary and the same signs in the continuous signing. On average, the dictionary signs are more than twice longer than signs from continuous signing. The average duration of signs in our dataset is 0.81/0.38 seconds in dictionary/continuous signing. There are also differences in signs that consist of repetitive moves. Usually, more repetitions are made in dictionary items than in continuous signing. Those differences are insignificant in human understanding of the sign but enlarge the measured distance.
The transitions are synthesized with a constant length and such an approximation does not correspond with the observed reality. The cubic spline interpolation is also heavily dependant on the annotation's precise selection of the start

and the end point and also does not respect the nature of the human movement.

9. Conclusion

We presented a new 3D motion capture dataset of Czech Sign Language (CSE), which we would like to share with the community. Its main purpose is to provide the data for further analysis and data-based automatic synthesis of CSE utterances. The dataset was recorded using the state-of-the-art motion capture technology to provide the most precise trajectories of the motion. The size of the dataset and the precision of tracked components are comparable to the best existing datasets for other SLs. The dataset contains trajectories of body, arms, hands and face markers recorded at once in order to provide consistent data without the need for the time alignment.

We introduced a baseline for the data-driven synthesis of SL utterances and suggested a method for objective data evaluation in the form of normalized DTW algorithm and Euclidean distance.

In future work, we will focus on improving the quality of the synthesis by using machine learning techniques and the normalized DTW distance as an objective function. We would also like to verify the correlation between objective and subjective evaluations.

We also would like to further improve synthesis by adding a non-manual property as well as other more complex SL grammar concepts. This will require annotations in more than one-tier. The additional annotation can be done in semi-automatic or fully automatic mode. It will also be beneficial to use multiple annotators on the same task to eliminate human errors and improve the precision of an annotation.

10. Acknowledgement

This work was supported by the Ministry of Education of the Czech Republic, project No. LTARF18017. This paper was supported by the Ministry of Education, Youth and Sports of the Czech Republic project No. LO1506. This work was supported by the European Regional Development Fund under the project AI&Reasoning (reg. no. CZ.02.1.01/0.0/0.0/15 003/0000466).

11. Bibliographical References

Berndt, D. J. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA.

Cao, Z., Simon, T., Wei, S., and Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1302–1310.

Crasborn, O. and Sloetjes, H. (2008). Enhanced ELAN functionality for sign language corpora. In *6th International Conference on Language Resources and Evaluation (LREC 2008) 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, pages 39–43.

Gibet, S. (2018). Building French Sign Language Motion Capture Corpora for Signing Avatars. In *Workshop on the Representation and Processing of Sign Languages: Involving the Language Community, LREC 2018*, Miyazaki, Japan, May.

Huenerfauth, M., Zhao, L., Gu, E., and Allbeck, J. (2008). Evaluation of american sign language generation by native asl signers. *ACM Transactions on Accessible Computing (TACCESS)*, 1(1):1–27.

Krňoul, Z., Kanis, J., Železnỳ, M., and Müller, L. (2016). Semiautomatic data glove calibration for sign language corpora building. In *7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining, LREC*.

Krňoul, Z., Jedlička, P., Kanis, J., and Železný, M. (2016). Toward sign language motion capture dataset building. In *Speech and Computer*, pages 706–713, Cham, 08. Springer International Publishing.

Lu, P. and Huenerfauth, M. (2010). Collecting a motion-capture corpus of american sign language for data-driven generation research. In *Proceedings of the NAACL HLT 2010 Workshop on Speech and Language Processing for Assistive Technologies*, pages 89–97. Association for Computational Linguistics.

McDonald, J., Wolfe, R., Wilbur, R. B., Moncrief, R., Malaia, E., Fujimoto, S., Baowidan, S., and Stec, J. (2016). A new tool to facilitate prosodic analysis of motion capture data and a data- driven technique for the improvement of avatar motion.

Naert, L., Larboulette, C., and Gibet, S. (2017). Coarticulation analysis for sign language synthesis. In Margherita Antona et al., editors, *Universal Access in Human–Computer Interaction. Designing Novel Interactions*, pages 55–75, Cham. Springer International Publishing.

Sedmidubsky, J., Elias, P., and Zezula, P. (2018). Effective and efficient similarity searching in motion capture data. *Multimedia Tools Appl.*, 77(10):12073–12094, May.

Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 107–112
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

A survey of Shading Techniques for Facial Deformations on Sign Language Avatars

Ronan Johnson, Rosalee Wolfe
DePaul University, Chicago, IL, USA
sjohn165@depaul.edu, wolfe@cs.depaul.edu

Abstract

Of the five phonemic parameters in sign language (handshape, location, palm orientation, movement and nonmanual expressions), the one that still poses the most challenges for effective avatar display is nonmanual signals. Facial nonmanual signals carry a rich combination of linguistic and pragmatic information, but current techniques have yet to portray these in a satisfactory manner. Due to the complexity of facial movements, additional considerations must be taken into account for rendering in real time. Of particular interest is the shading areas of facial deformations to improve legibility. In contrast to more physically-based, compute-intensive techniques that more closely mimic nature, we propose using a simple, classic, Phong illumination model with a dynamically modified layered texture. To localize and control the desired shading, we utilize an opacity channel within the texture. The new approach, when applied to our avatar "Paula", results in much quicker render times than more sophisticated, computationally intensive techniques.

Keywords: Sign Language Synthesis, Nonmanuals, Illumination Models, Avatars

1. Introduction

In all humans, facial movements can convey important information including emotion, social cues, and a person's general demeanor. However, in signed languages, such facial movements are also used to convey important linguistic information. Notable examples include differences in eyebrow position when asking yes/no versus "wh" questions, and nonmanual adjectives and adverbs that co-occur with their corresponding sign (Baker-Shenk, 1985; Reilly et al., 1990). Although a hearing person might initially try to understand sign language by following the movement of the hands, in fact, native signers focus primarily on a user's face (Siple, 1978). This is such an important component of comprehension that when viewing an avatar with displaying insufficient facial distinction, many signers become irritated or distracted. This can decrease their understanding of the utterances being portrayed (Kipp et al., 2011). To this end, accurate facial movements, clearly portrayed are of utmost importance when recreating signed language through an avatar.

Visual legibility is crucial to such portrayals. Subtle nuance and co-occurring actions in facial movements communicate important linguistic information. For example, the intensity of a nonmanual modifier to a verb corresponds to the intensity of the modification. Additionally, puffed cheeks in conjunction with a sign conveying a large object will indicate an extreme size difference. In English, we might interpret this as the difference between "large" and "huge" (Baker-Shenk, 1983).

Using avatars for linguistic research with Deaf participants requires making the movements and visual look of real life signers as closely as possible. However, one of the many challenges in rendering sign language avatars is properly gauging the trade-off between realism and computational efficiency. While it is important that digital visualizations mimic real life signers as closely as possible, one must also be cognizant of the technical challenges involved with rendering complicated avatars in real-time.

One such consideration is the lighting conditions used to illuminate an avatar. It has been observed that Deaf participants respond more positively to lighting conditions that clearly illuminate the hands and face. Furthermore, light and shadows that improve the realistic appearance of an avatar makes the avatar seem more like an interpreter (Kipp et al., 2011).

This paper presents a discussion of three primary illumination models in computer graphics and their potential application to realistic portrayal of sign language avatars. We conclude with an improved approach involving an implementation of layered textures to achieve a solid solution to the complexity/realism trade-off.

2. Illumination Models

In real life, the way light interacts with objects can be taken for granted. The same cannot be said for the world of computer graphics. In order for a computer generated object to be rendered visible on a screen, the computer must calculate the color of each pixel based on the conditions present in the 3D scene. Such conditions include the presence of any geometry, available light sources, and the surface properties of the geometry which determine how light will interact with any given object. Illumination models are algorithms designed to calculate light reflection from objects in a scene (Hall, 1986). There are three main types of illumination models, each with their own benefits and drawbacks. They are listed as follows:

- Local illumination
- Global illumination
- Semi-global illumination

2.1. Local Illumination – The Phong Illumination Model

Local illumination models empirically calculate the color of a point on a surface based on the properties of that surface and the properties of any light coming directly from a

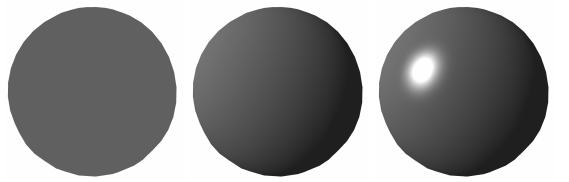


Figure 1: The ambient, diffuse, and specular components of the Phong illumination model.

light source that strikes the point (Phong, 1975). This set of algorithms is easy to compute, often requiring a minimal amount of coding. However, these models often prove insufficient for truly realistic displays of geometry, as they only a rough approximation of the natural effects of the light.

One of the most fundamental local illumination models is known as the Phong illumination model, named after its initial proposer Bui Tuong Phong in 1975. This algorithm computes the color of a point on an object by breaking the surface properties out into three components: the ambient light being reflected around the scene, the diffuse reflection of a direct light source, and the specular reflection from a direct light source. The algorithm then computes the sum of these components for each of the red, green, and blue color channels for each light present in the scene (Phong, 1975). This summation is the final result rendered to the screen. A breakdown of each of these components is shown in Figure 1.

Because local illumination only considers light coming directly from a light source, and ignores light from reflecting surfaces, the resulting shading can be very harsh and unnatural. To counteract this effect, illumination models add a constant lighting term called ambient light. It represents an amount of light present on an object in the absence of any direct light source.

In a local illumination model, diffuse reflection is the even scattering of a direct light according to Lambert's law. The strength of this reflection only depends on the strength of the incident light and the surface's light absorption properties. The viewing angle does not affect it (Cohen and Greenberg, 1985). The specular reflection of a surface is the amount of light reflected from the surface, based on the direction at which the object is viewed (Phong, 1975). It is responsible for the highlight or "shiny spot" seen in highly polished surfaces.

While this illumination model is simple to compute, it has several major drawbacks. For one, it is unable to render shadows cast by one object onto another object. We can observe this by placing another piece of geometry directly below the sphere in Figure 2. Although the sphere is physically touching the floor, the absence of shadow makes the sphere appear to float above the surface. We also see that this model also does not compute the effects of indirect light or bounced light, remaining reliant on direct light sources. In general, local illumination models such as Phong are unable to render the kind of complex detail we need for maximum legibility in sign language production. In order to achieve these properties, we must turn to more advanced illumination models.

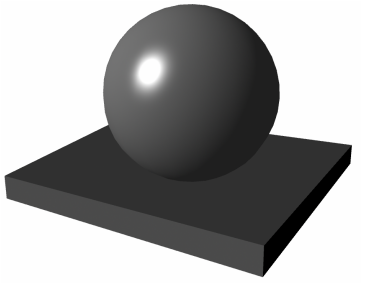


Figure 2: Phong illumination does not calculate cast shadows or indirect light.

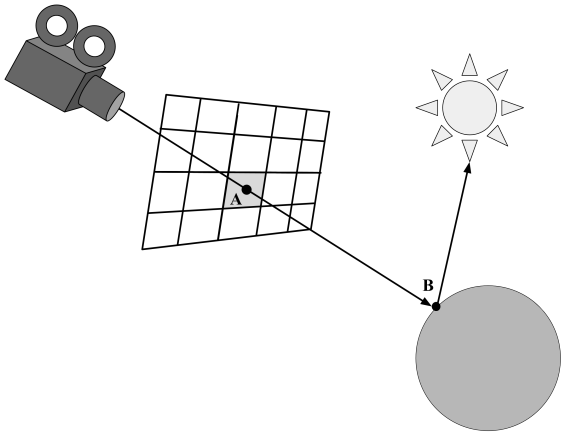


Figure 3: The calculation of the color of a pixel using ray tracing.

2.2. Global Illumination

2.2.1. Ray Tracing

Ray tracing was actually first described in 1532 by Albrecht Dürer, a renaissance artist who famously learned to draw by looking at his subjects through a grid (Hofmann, 1990). First described by Appel in 1967, this illumination technique uses the physical properties of light to simulate how photons would interact with a scene (Appel, 1967).

In the real world, light is emitted from a light source, then it bounces around a room until some of its photons reach a viewer's eye. This allows the viewer to see the object. Understandably, some photons may take more bounces than others. Others may never reach the viewer at all. It is a waste of processing time to compute light bounces that will never make it to a viewer. To avoid processing light that never enters the viewer's eye, ray tracing computes the light bounces backwards. A ray is begun at the center of each pixel in the computer generated camera's view. The algorithm then traces this ray across the scene until it intersects with some object. A final ray is then cast from this point towards the light source. The final color of the pixel is the result of the light and object properties that the ray encounters (Appel, 1967).

In 1979, this technique was expanded by Turner Whitted. His algorithm generated multiple new rays at the points of

intersection, which could then bounce across the scene until they either hit a light source, or some maximum number of bounces was reached. The final color of the pixel is then back-computed from the combined information collected by each generated ray. In doing so, ray tracing became capable of yielding photorealistic images including cast shadows, reflections from shiny surfaces, and refraction through water and glass (Whitted, 2005). However, ray tracing's main drawback is its computational demands. At present, real-time ray tracing is a heated area of research, especially in the field of video games. The most successful techniques at time of writing are only achievable with dedicated graphics hardware and specialized APIs (Liu et al., 2019).

2.2.2. Radiosity

Of the illumination models surveyed, radiosity most closely simulates the actual physics of lighting. Radiosity, a form of global illumination, is similar to ray tracing, except that it tracks all the rays starting at the light source(s) and bounces them around the entire scene. As a white light ray bounces off a surface, it can change color, depending on the surface's properties. It treats the diffuse reflection of nearby objects as an additional ambient light source to simulate indirect lighting. Because of this, radiosity allows color bleeding from nearby objects to be visible. Because the ambient illumination is no longer a constant term, this technique is capable of rendering fine detail and soft, highly realistic shadows that deepen in corners (Cook and Torrance, 1981). This effect can be observed in Figure 5.

Of course, this comes at the cost of render time. Unlike ray tracing, radiosity is view-independent, meaning the lighting of the whole scene is computed, not just the light as seen from the camera (Cohen and Greenberg, 1985). Understandably, this would greatly increase the complexity of the computations. Compare this to ray tracing which reduces the complexity by calculating the illumination for only the geometry visible to the camera.

Figure 4 shows a diagram of the basic action of the light rays during radiosity calculations. The illumination of a given point C is calculated based on the sum of the illumination components of the rays leading from the light source to that point. Point A is where the direct light ray initially encounters geometry. The illumination at point B will then be calculated taking into consideration the properties of point A. Point C will be calculated with the properties of rays AB and BC.

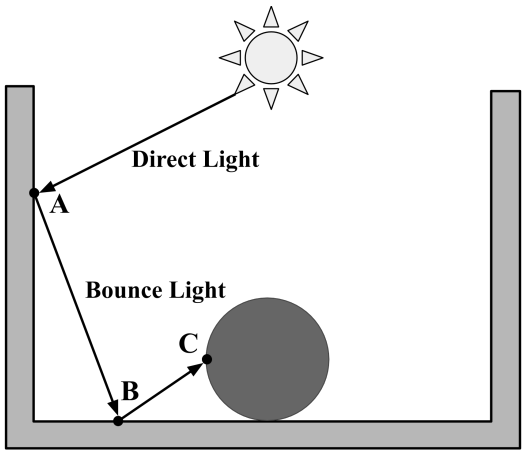


Figure 4: The calculation of radiosity.

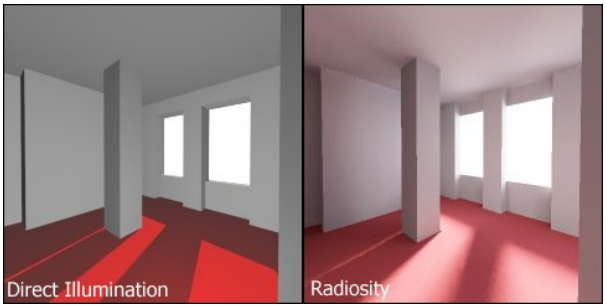


Figure 5: The shadows created by radiosity are softer and more realistic than direct illumination through ray tracing (Elias, 2006).

2.3. Semi-Global with Ambient Occlusion

Semi-global methods attempt to produce the effects of global illumination but at a lower computational cost. For highly dynamic scenes, real-time radiosity is still untenable. However, it is possible to achieve some amount of the effect using approximations (Ritschel et al., 2009). One such approximation is ambient occlusion. This refers to the effect adjacent geometry has in blocking some sources of bounce light from reaching a specific point (Scherson and Caspary, 1987). True ambient occlusion is a by-product of radiosity, but it can be approximated using a variant of ray tracing and added on later using a render pass (Miller, 1994). The classic example is the way the corners of a room appear darker than the adjacent wall, as described in Figure 6.

Ambient occlusion simplifies the calculations necessary for creating the soft shadows of radiosity. This greatly improves performance in real-time applications while providing significant improvements in image quality (Ritschel et

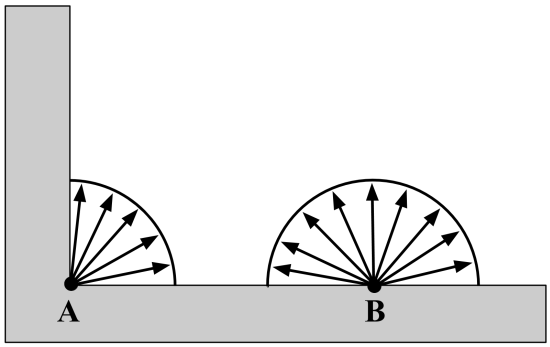


Figure 6: Point B can receive bounce light from any point in the scene. Point A can receive bounce light from fewer sources. This causes the point to be darker.

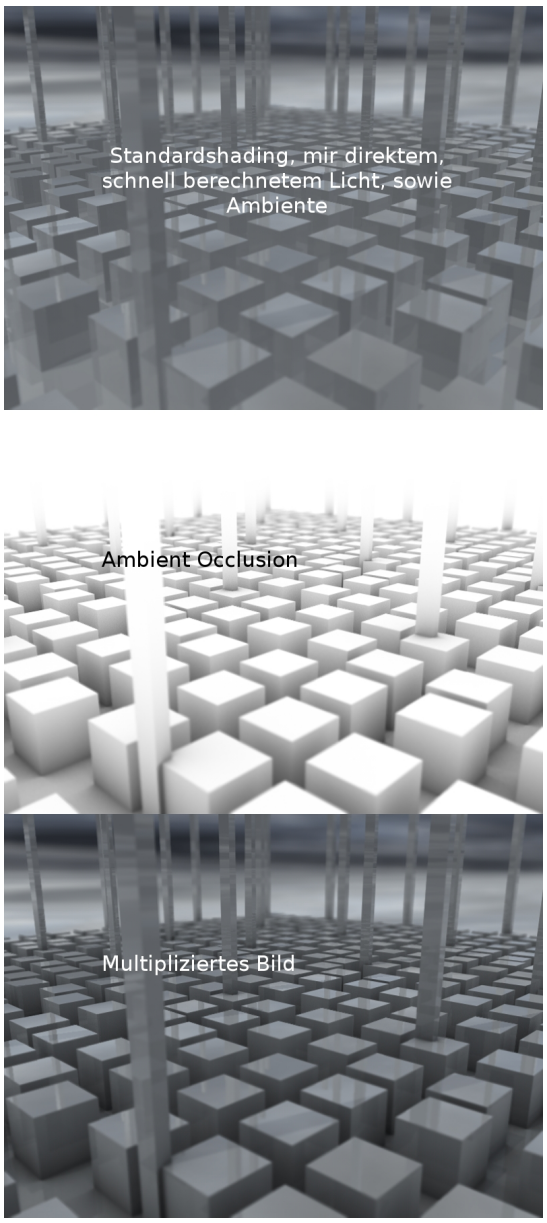


Figure 7: The addition of ambient occlusion heightens the realism of computer generated images (TheWusa, 2006).

al., 2009).

The sequence of images in Figure 7 shows the level of realism that can be achieved with ambient occlusion. The top image shows a scene rendered without ambient occlusion. The middle image shows the incident illumination computed by an ambient occlusion pass as an approximation of radiosity. The bottom image shows the results of combining this ambient occlusion with the initial rendering pass.

3. Trade-offs Among the Three Approaches

The Phong illumination model is easy to compute and yields fast results. However, the final renders often lack the amount of desired realism. In particular, it is incapable of rendering shadows. Ray tracing can compute these shadows, but they remain unrealistically harsh. Although soft shadows are possible with ray tracing, it comes at the cost of greatly increased render times (Scherson and Caspary, 1987). Radiosity provides the greatest level of realism, but the the most computationally expensive of all the options. The benefits of the soft shadows can be approximated with ambient occlusion. This is faster than true radiosity, it still represents a huge time cost. Optimizing ambient occlusion for real-time applications remains an ongoing area of research (Jiménez et al., 2016).

4. Applications to Sign Language Avatars

We require that our avatar be able to closely mimic reality as much as possible. To that end, the global and semi-global illumination models provide us the best opportunity for realistic rendering. However, their complexity makes them untenable for our real-time applications. We therefore turn to the Phong illumination model, which provides us the render times we need to implement our software on a variety of computers. In order to overcome the limitations in the level of detail this model can portray, we add pre-rendered texture maps to create shaded areas of fine detail. This gives us the speed of local illumination combined with the realism of radiosity.

4.1. A Layered Texture Approach

Previous work in applying the layered texture technique involved controlling dynamic textures to indicate facial deformations. There are several instances in portraying facial deformations on sign language avatars where the presence of shadows is necessary to clearly convey an appropriate amount of detail. One major example is the horizontal wrinkles that form on the forehead when the eyebrows are raised (Wolfe et al., 2011). Creating physical wrinkles requires the addition of significantly more geometry in that area. Our previous work in optimizing our avatar, “Paula” for real-time rendering describes the challenges associated with rendering large amounts of geometry (McDonald et al., 2016). In order to avoid major increases in render times, we elected to implement a layered texture that would fade in and out depending on the position of the eyebrows.

A further exploration of the semi-global illumination model of fast ambient occlusion yielded an insight into the problem of creating a more realistic, clearer face rendering that emphasized all necessary facial poses while achieving video rates while rendering. This approach pre-renders the shadows that appear in folds of the skin when a person extends or contracts sets of facial muscles. Additionally, the shading will become darker or lighter depending on the intensity of the deformation. For example, the higher the eyebrows are raised, the darker the forehead wrinkles become.

One hindrance of extending the previous approach was the opaque nature of the textures. Painting the wrinkle shadow directly onto the texture of the face and then turning that texture on and off precludes the introduction of additional wrinkles, which are necessary for producing co-occurring facial actions. In the previous approach it was possible to show the forehead wrinkles or enhanced eye creases, but to

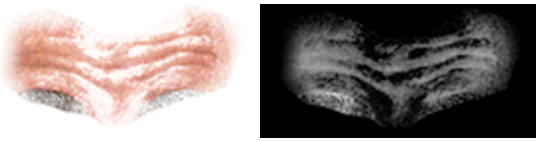


Figure 8: The wrinkled brows texture is isolated on the left. The right shows the alpha channel for this texture.

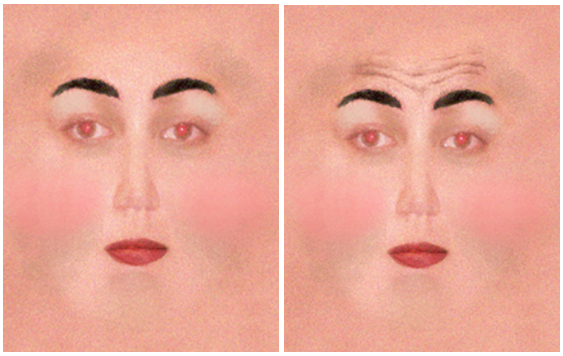


Figure 9: The left image is the base texture for our avatar's face. The right image shows the results of layering the wrinkled brows texture over this base using its alpha channel.

show both at the same time required a prohibitively complex masking function.

The question then was how to combine multiple pre-rendered textures. In addition to the classic red, blue, and green channels of an image, a new approach introduces an alpha channel in the pre-rendered textures. An alpha channel controls the visibility of the texture. The color black denotes the areas of an image that are transparent, as shown in Figure 8. Therefore, when layered atop one another, the base texture remains unchanged while the additional layers are able to be "turned on and off" individually. Figure 9 shows the results of this layering. This provides complete control over selectively adding simulated ambient occlusion to localized areas. This approach removes the limitation of using only one texture map at a time. Not only are the render times unaffected by this addition, but it also allows greater freedom and flexibility in the final look of the renders. This is much faster than attempting to physically model and shade the smaller folds of the fast. Changes made by artists can be seen almost immediately.

This technique is also useful for facial areas where the geometry deformations do exist, but a conventional illumination model is not producing the level of shadows necessary for effective portrayal. The primary deformation that motivated this work was in the cheek puff action where an amount of air is pushed into the cheek, creating a rounded, protruding shape. However, the previous illumination model and lighting setup were inadequately rendering the shadows necessary to clearly portray the deformation. Figure 10 shows the original cheek puff deformation on the left and the same deformation with an added texture layer creating the ambient occlusion effect. With this alone, there is a perceivable increase in legibility of the

Figure 10: Paula displaying the puffing action with (bottom) and without (top) additional ambient occlusion.

cheek puffing action.

As with the eyebrows, the intensity of the layered textures can be dynamically controlled with the intensity of the deformation. This creates a fading effect as the face animates so there is no distraction from the texture suddenly popping in and out. This is also important for portraying subtleties of the facial deformations denoting linguistic information. With variable intensity, Paula is capable of portraying the full variety of intensity modifiers to concurrent manual signs.

We are able to successfully implement this approximation based on two assumptions. The first assumption is the use of a static lighting setup, meaning there are limited changes in the ambient light as Paula's head turns relative to the viewer. Because there is limited motion relative to the lighting, the shadows remain consistent and predictable. The other assumption is the predictability of Paula's behavior. The location and intensity of the cheek puff deformation is within a strict range of parameters. This means we do not have to account for the general case of shading every possible position and deformation.

5. Conclusion and Future Work

A new technique for emphasizing facial deformations through texture mapping yields greater legibility while still maintaining the superior performance of local (simple) illumination models when rendering. This technique will be applicable to other areas of the face that experience similar deformations. One such example is the upper area of the cheek near the eyelids that pushes up and forward when expressing a smile or scrunching the face. Future work includes creating additional texture layers and continuing to compare render times with a conventional single-texture model to assess whether there is a maximum practical limit to the number of layers. This will maintain Paula's ability to be run on machines without high-end processors and graphics cards.

6. Acknowledgment

We would like to acknowledge Elena Jahn for her work on the makeup textures for Paula.

Bibliographical References

Appel, A. (1967). The notion of quantitative invisibility and the machine rendering of solids. In *Proceedings of the 1967 22nd national conference*, pages 387–393.

Baker-Shenk, C. (1983). A microanalysis of the nonmanual components of questions in american sign language.

Baker-Shenk, C. (1985). The facial behavior of deaf signers: Evidence of a complex language. *American Annals of the Deaf*, 130(4):297–304.

Cohen, M. F. and Greenberg, D. P. (1985). The hemi-cube: A radiosity solution for complex environments. *ACM Siggraph Computer Graphics*, 19(3):31–40.

Cook, R. L. and Torrance, K. E. (1981). A reflectance model for computer graphics. *ACM Siggraph Computer Graphics*, 15(3):307–316.

Elias, H. (2006). https://upload.wikimedia.org/wikipedia/commons/5/55/radiosity_comparison.jpg. licensed under the gnu free documentation license, version 1.2.

Hall, R. (1986). A characterization of illumination models and shading techniques. *The Visual Computer*, 2(5):268–277.

Hofmann, G. R. (1990). Who invented ray tracing? *The Visual Computer*, 6(3):120–124.

Jiménez, J., Wu, X., Pesce, A., and Jarabo, A. (2016). Practical real-time strategies for accurate indirect occlusion. *SIGGRAPH 2016 Courses: Physically Based Shading in Theory and Practice*.

Kipp, M., Heloir, A., and Nguyen, Q. (2011). Sign language avatars: Animation and comprehensibility. In *International Workshop on Intelligent Virtual Agents*, pages 113–126. Springer.

Liu, E., Llamas, I., Kelly, P., et al. (2019). Cinematic rendering in ue4 with real-time ray tracing and denoising. In *Ray Tracing Gems*, pages 289–319. Springer.

McDonald, J., Wolfe, R., Schnepp, J., Hochgesang, J., Jamrozik, D. G., Stumbo, M., Berke, L., Bialek, M., and Thomas, F. (2016). An automated technique for real-time production of lifelike animations of american sign language. *Universal Access in the Information Society*, 15(4):551–566.

Miller, G. (1994). Efficient algorithms for local and global accessibility shading. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, pages 319–326.

Phong, B. T. (1975). Illumination for computer generated pictures. *Communications of the ACM*, 18(6):311–317.

Reilly, J. S., McIntire, M. L., and Bellugi, U. (1990). Faces: The relationship between language and affect. In *From gesture to language in hearing and deaf children*, pages 128–141. Springer.

Ritschel, T., Grosch, T., and Seidel, H.-P. (2009). Approximating dynamic global illumination in image space. In *Proceedings of the 2009 symposium on Interactive 3D graphics and games*, pages 75–82.

Scherson, I. D. and Caspary, E. (1987). Data structures and the time complexity of ray tracing. *The Visual Computer*, 3(4):201–213.

Siple, P. (1978). Visual constraints for sign language communication. *Sign Language Studies*, (19):95–110.

TheWusa. (2006). https://upload.wikimedia.org/wikipedia/commons/9/91/ambientocclusion_german.jpg licensed under the gnu free documentation license, version 1.2.

Whitted, T. (2005). An improved illumination model for shaded display. In *ACM Siggraph 2005 Courses*, pages 4–es.

Wolfe, R., Cook, P., McDonald, J. C., and Schnepp, J. (2011). Linguistics as structure in computer animation: Toward a more effective synthesis of brow motion in american sign language. *Sign Language & Linguistics*, 14(1):179–199.

Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 113–116
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

Use cases for a Sign Language Concordancer

Marion Kaczmarek, Michael Filhol
Université Paris Saclay, CNRS, LIMSI
Orsay, France
kaczmarek@limsi.fr, michael.filhol@limsi.fr

Abstract

This article treats about a Sign Language concordancer. In the past years, the need for content translated into Sign Language has been growing, and is still growing nowadays. Yet, unlike their text-to-text counterparts, Sign Language translators are not equipped with computer-assisted translation software. As we aim to provide them with such software, we explore the possibilities offered by a first tool: a Sign Language concordancer. It includes designing an alignments data base as well as a search function to browse it. Testing sessions with professionals highlight relevant use cases for their professional practices. It can either comfort the translator when the results are identical, or show the importance of context when the results are different for a same expression. This concordancer is available online, and aim to be a collaborative tool. Though our current data base is small, we hope for translators to invest themselves and help us to keep it expanding.

Keywords: Sign Language, Concordancer, CAT

1. Introduction

Translation is part of our world, and Sign Languages (SL) should be no exception. As far as France is concerned, the Law for Equal Rights and Opportunities, Participation and Citizenship of Persons with Disabilities published in 2005 recognizes French Sign Language (LSF) as a fully-pledged language, and as a « language of the Republic, the same way as French ». This law means that any public place must be able to welcome deaf people (either by forming their staff to SL or by means of professional SL interpreters), but also every piece of information provided (videos, written documents, or audio announcements). Helped by the CRPD in 2008 which puts emphasis on the right of people with disabilities to fully access information, the need for SL translated content is still growing. To try and fit the needs, a master degree in French/French Sign Language translation and mediation was created in 2011. However, there are still very few professional sign language translators.

And those few translators are not equipped with tools as the other translators can be. Indeed, no current Computer Assisted Translation (CAT) software is able to support SL translation.

Our aim is to provide CAT software dedicated to SL translation. The concordancer referred to in this article is a part of a bigger project. This paper, and our previous studies, are based on French Sign Language (LSF) as working language.

2. Brief state of the art

As our goal is to specify CAT software for SL translation, we first needed to learn more about it. To do so, we conducted studies involving professional SL translators and interpreters, including brainstorming sessions and observing them at work to analyse their practices. (Kaczmarek & Filhol, 2019). The results highlight their needs and the most common problems encountered, such as the scarcity of SL resources, the time spent looking for them as they are not always well referenced, the need for context and encyclopedic knowledge… Most of the identified steps taken in the process of text-to-Sign translation could benefit from already existing tools if they were able to support SL. We also pointed out what the major differences are between text-to-text translation and text-to-sign translation, sorted in four categories: no written form for SL, a *principle of linearity*, need for encylcopedic knowledge and CAT tools adaptation issues arising from the previous categories. (Kaczmarek & Filhol, 2019)

On the other hand, the major innovation brought by CAT software is Translation Memory (TM). It allows the translator to store prior work and reuse it later. Once a segment of source text is translated, the source-translation pair is stored in memory. When the translator encounters a similar one, the TM automatically suggests the prior translation. It can be shared with colleagues or even provided by the client himself. The is a time saving tool which has had a great impact on the everyday practices of translators, evolving from to translating from scratch to mostly post-editing TM entries and suggestions. (Lagoudaki 2006; O'Hagan 2009).

A concordancer is, regardless of the languages, a search engine which can look through corpora and list each and every occurrences of a queried word. When it comes to translation, bilingual concordancers are used. The query is done in a source language, and results are provided with an aligned translation in target language, in our case LSF. Such tool allows translators to look up words or expressions in context, to determine how common they might be or with which style of discourse they are more often associated with.

The Danish Sign Language Dictionary[1] (Kristoffersen & Troelsgard, 2012) includes a concordancer view in the results' display. Each sign is given its definition and used in a example sentence, in context. However, this concordancer view depends on the dictionary and only counts one example per sign/meaning. The iLex view of

[1] http://www.tegnsprog.dk/

the DGS korpus also includes some concordance tokens, to work on meanings and sub-meanings from a dictionary point of view (Langer, Müller & Wähl, 2018).

In our case, we want a concordancer to work as a stand-alone tool, and focused on the variety of the examples rather than on covering each sign. The focus is not on the meaning, but on the impact of context on the translation choices. The next part treats about the creation of our own data base. The next section explores how we did.

3. Designing a concordancer

SLs do not have editable written forms, so video is the most common way to keep trace of it. This brings a problem when it comes to the adaptation of a TM tool for SL, as chaining video extracts from previous translations alone would result in an unacceptable translation.

As previously said, TM stores alignments, in other words pairs composed of two text segments, where one is the the translation of the other for two given languages. This data can be searched with concordancer. This is why we elaborated a SL concordancer to keep the benefits of the TM. The alignments consist in pairs made of a segment of the source text, and its SL translation identified in a video with time tags. Such alignments are stored in a data base built by the users themselves, and which can be shared just like a TM.

3.1 Alignments data base

The creation of such data base is the main topic of another article (Kaczmarek & Filhol, 2020) in which you can find more details. The next paragraph explains briefly its key points.

As there is currently no automatic way to produce text-SL video alignments, we built the first data base ourselves by aligning manually. We used a French–LSF parallel corpus of forty short news texts, of three to five lines each in a journalistic style ("40 brèves"[2]). Each text was translated by three different professional SL translators, and the resulting translations filmed using two cameras for a front and a side view, for a total of 120 videos of an average 30-seconds duration. We used it because it is a parallel corpus of short translations , so that each video already provides us with a useful alignment (the 30-seconds video can be aligned with the 2-3 lines text). As the videos are shorts, those are still interesting alignments to include in the data base.

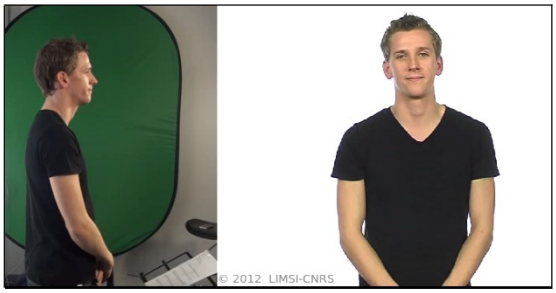


Figure 1 : A screenshot of the video set-up

[2] https://www.ortolang.fr/market/corpora/40-breves

We chose a few smaller segments : words, idiomatic expressions, grammatical phenomena or figure of speech. As we had three different signers, we were free to vary the spans for a given expression. The identified expressions were also chosen either based on their lack of standard signs, or on variety of translations proposed by the signers. For each text segment identified, we search for its translation in the associated video and extract the corresponding time-tags. The selected segments are also suitable for aligning in terms of simultaneity. We cannot for example, align an adverb in the text with only the facial expression of the signer.

Alignments are stored in a data base with the following format:

<TxtID, start pos., length, VidID, start time, duration>

- TxtID is the identification code of the text, ranging from 1A to 1T and from 2A to 2T. This code allows to retrieve them in their own storage space.
- Start pos. is the position of the first character from the text segment, in the source text.
- Length is the number of characters in the segment
- VidID is the unique identification of the video in their own storage space, which allows to retrieve them.
- Start time is the time tag corresponding to the beginning of the SL translation in the video.
- Duration is the total duration of the segment's translation.

3.2 Search function

The search algorithm looks for matches for the string entered as a query. A query is usually an exact match, but can include wildcards such as:

- A "#" suffix to match any word ending, with or without a limit in length. For example, Europ# will match European, Europeans, Europe…
- A "##" suffix to match word sequences which allows to split queries. For example: bring ## forward will match bring something forward, bring forward, or bring the solution forward.

If the expression queried by the user has been previously aligned, the concordancer answers with the smallest span found. If it has not been aligned but still previously translated, the concordancer answer with the video of the entire source text that contains the query. If it has never been translated before, the concordancer cannot answer the query.

The concordancer itself is currently available on-line, at the following address: platform.postlab.fr . You can either test it using a public test account, or create your

personal account on the website. In either case, contact us via e-mail to: kaczmarek@limsi.fr so we can provide you with the information needed.
If you are interested to contribute to the data base, please take contact with us. We are currently collecting feedback from our users, which may lead to a later communication.

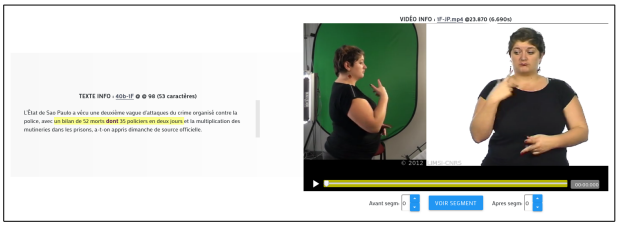


Figure 3 : The result page format

The figure above is a screen shot of the concordancer, showing one of the items matched for the query *dont* which means "whose", "of which" or "including". On the left is the entire source text in which the query was matched. The segment aligned appears in yellow, and the exact query is in bold. The line above the text gives information about it: the TxtId, start pos. and length of the segment, as well as a link to open the entire text in a new tab. On the right, the video appears centered on the segment (which also appears in yellow on the time line). The title above provides the same kind of information as for the text, and a link to open the entire video in a new tab. The video is looping on the segment, and the buttons below allow the user to add left or right context (in seconds) around the segment.

As mentioned by the professional translators and interpreters during our prior studies, LSF resources are rare. And those rare resources are often badly documented or sometimes hard to access. In addition to the convenience that such tool can bring to the translators' everyday practices, it can also be an opportunity to explore the language in a unique way. To an extent, this tool could have a certain use in teaching not only LSF, but also in teaching translation and interpretation methods by displaying in a very readable way a list of examples, counter-examples, as well as unique constructs to think about in class. Following are 3 examples of phenomena we did observe while working on our data base, which could help either SL learners or SL linguists to understand better the language, to speak it or describe it. Those are also typical use cases for a text-to-Sign translator. During a previous workshop, we presented some professional translators with texts to translate into Sign Language. We built those texts around idiomatic French expressions or complex semantic or syntactic rules. When asked which part of the text they would most likely search for in a concordancer, those three examples were among the most identified ones.

4. Use cases

4.1 No standard sign, yet common form

Some frozen French expressions do not have standard signed equivalent. It is the case for the French expression *fin de non recevoir*, which can be translated in English by "refusal to consider one's request". But the fact that there is no standard way of signing it does not mean that it is untranslatable. When searching the concordancer for *fin de non recevoir*, three results came up. The three translators worked alone on their translation, still they chose a similar construction to translate this expression. The sign used here is the one for "to reject". Their facial expressions are also similar, as well as the spacial construct they are using. Overall, the entire source text is translated in a quite similar way, meaning the context has only a low impact on the translation choices.

Figure 4 : Three ways to translate *fin de non recevoir*.

Three times the same construction seems trustworthy even if there is no standard sign. This kind of results can comfort the translator, either in his choice of translation, or encourage him to reuse the same construction in his own work.

4.2 No standard sign, and no common form

Here again, the French expression *mis à mal*, which means "suffering from a negative effect of something or someone", or "to be harmed", does not have an out-of-context equivalent is SL. Still, we can find three examples of its translation in our data base, and the three of them are different. The sentence here was "Hopes for peace si Sri Lanka are once again dashed after a major military offensive against Tamil rebellion".

Figure 5: Three ways to translate *mis à mal.*

On figure 5, the signer on the left uses the sign for "break", and the one on the right the sign for "difficult". The signer in the middle uses an iconic structure based on the French sign for "hope", which he signs falling down crumbling. Here, the way of translating is more influenced by the context than in our first example.

Three different results for the same expression translated from the same source text. This kind of results allows us to see how different matches can be in context. This is

very useful for professional translators to build on what has already been done, but also for the learners to better understand the finer points of the language.

4.3 Cause/effect relationship

The concordancer is also a interesting way to observe grammatical phenomena such as this one. Cause/effect relationships can be translated in many ways depending on two things: the translator's choice and the context. The translator is free to use the sign for "then", as the signer on the right does on figure 6. The two ohters made the choice of using iconic structures to depict the event: an underwater earthquake occurs and causes a tsunami (*un tsunami causé par un séisme sous-marin* in French). For the left and middle signers, the cause/effect relationship lies in the order of events and in the facial expressions, as well as in the dynamics of their speech. There is a specific transition time between the two events mentioned, and their signs.

Figure 6: The ways of translating *causé par*.

Figure 7: Translation for *provoqué par*.

On Figure 7, the sentence to translate was "a landslide caused by heavy rains". Both of them picture the waterlogged ground, and then the landslide itself. The screen shots were taken right before the landslide part, and we can see that the cause/effect relationship relies here again on the timeline of the vents but also in the facial expressions, and the very short transition time between the two, with raised eyebrows and chin as if to call for the viewer's attention.

5. Conclusion

Designing a SL concordancer first implied to build an alignments data base. We are aware that our first data base is rather small, but hopefully it will keep expanding. The examples detailed comfort the relevance of a SL concordancer as a tool to equip the translators. The very first feedback we received showed enthusiasm and interest in our work. The first prototype for the concordancer is fully working, and we are now waiting for some more specific feedback about the function itself, in an iterative process to converge on the most adequate kind of tool for them to use in their everyday practices.

We are now working on a alignment function, which would allow our users to create their own alignments data base in an easy way. Based on pairs of text and signed video displayed alongside (where one is the translation of the other), the user can select a segment in the text and identify the corresponding part in the video using start-stop tags. The alignment created this way is then stored in the user's data base, which is available in the search function. He can later consult his work, and report any problems or needs for data bases modification.

We hope that the translators who helped us during our studies will continue to invest themselves in this project, but also that others will join.

6. Bibliographical References

Kaczmarek, M. and Filhol, M. (2019). Assisting Sign Language Translation: what interface given the lack of written form and spatial grammar? In *proceedings of Translating and the Computer 41, London 2019*, p. 83-93.

Kaczmarek, M. and Filhol, M. (2020). Elaborating an Alignments Database for a Sign Language Concordancer, to be published in proceedings of LREC 2020.

Kristoffersen, J.H. and Troelsgård, T. (2012). The electronic lexicographical treatment of sign languages: The Danish Sign Language Dictionary. In *S. Granger and M. Paquot (Eds), Electronic Lexicography*, Oxford University Press, p. 293-318.

Lagoudaki, E. (2006). Translation Memory Survey 2006 : users' perceptions around TM use, in p*roceedings of Translating and the Computer 28, London 2006*.

Langer, G., Müller, A. and Wähl, S. (2018). Queries and Views in Ilex to support Corpus-based Lexicographic Work on German Sign Language (DGS), in *proceedings of LREC Sign Language Workshop 2018.*

O'Hagan, M. (2009). Computer-aided Translation (CAT) in *Baker, Mona/Saldanha, Gabriela (eds), Routledge Encyclopedia of Translation Studies*, London and New York, Routledge, p. 48-51.

UN General Assembly, (2007). *Convention on the Rights of Persons with Disabilities : resolution* / adopted by the General Assembly, 24 January 2007, A/RES/61/106.

Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 117–122
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

Towards Kurdish Text to Sign Translation

Zina Kamal, Hossein Hassani
University of Kurdistan Hewlêr, University of Kurdistan Hewlêr
Kurdistan Region - Iraq, Kurdistan Region - Iraq
{z.kamal3, hosseinh}@ukh.edu.krd

Abstract

The resources and technologies for sign language processing of resourceful languages are emerging, while the low-resource languages are falling behind. Kurdish is a multi-dialect language, and it is considered a low-resource language. It is spoken by approximately 30 million people in several countries, which denotes that it has a large community with hearing-impairments as well. This paper reports on a project which aims to develop the necessary data and tools to process the sign language for Sorani as one of the spoken Kurdish dialects. We present the results of developing a dataset in HamNoSys and its corresponding SiGML form for the Kurdish Sign lexicon. We use this dataset to implement a sign-supported Kurdish tool to check the accuracy of the sign lexicon. We tested the tool by presenting it to hearing-impaired individuals. The experiment showed that 100% of the translated letters were understandable by a hearing-impaired person. The percentages were 65% for isolated words, and approximately 30% for the words in sentences. The data is publicly available at `https://github.com/KurdishBLARK/KurdishSignLanguage` for non-commercial use under the CC BY-NC-SA 4.0 licence.

Keywords: Sign Language, Kurdish Language Processing, Kurdish to Sign, HamNoSys, SiGML

1. Introduction

The studies on sign language processing have been emerging, but many areas are still unexplored (Cormier et al., 2019). As might be expected, this area of research has even not been initiated yet for many under-resourced languages.

Kurdish, a multi-dialect language which is spoken by approximately 30 million people in different countries, is considered an under-resourced language (Hassani, 2018). It is also written in different scripts. The usage of the scripts changes according to the geographical situation (Hassani and Medjedovic, 2016).

The current literature does not report on visible research on Kurdish Sign Language (KuSL) processing, nor are there any publicly available resources for this topic. This research focuses on text to sign conversion for the Sorani dialect of Kurdish.

Sign language is the main communication method among the hearing-impaired community. This language is based on visual interaction rather than using sound. The interactions happen by manual and non-manual signs and finger spelling (Cooper et al., 2011). Hand and body movement, shape, orientation and location are within manual signs (Kelly et al., 2009), while facial expressions, eye gaze, and shoulder movement are called non-manual signs (Halawani, 2008). Furthermore, the finger spelling is used to spell letters of certain words, for example, names and technical terms that do not have sign equivalents (Liwicki and Everingham, 2009).

Normally, the communication between two hearing-impaired persons is smooth and understandable. The real challenge begins when a hearing person wants to interact with a hearing-impaired person (Wazalwar and Shrawankar, 2017).

Generally, if the target hearing-impaired person is educated, they try to communicate by exchanging written texts. Otherwise, they turn to a human sign language interpreter as a recourse if available, or else perhaps they end up with serious miscommunication (Wazalwar and Shrawankar, 2017).

Although the spoken Kurdish dialects use different lexicons (Ahmadi et al., 2019), the Kurdish Sign language, which is used in the Kurdistan Region of Iraq (KRI), uses the same lexicon among the hearing-impaired community regardless of the spoken dialect. While according to Jepsen et al. (2015) KuSL is not standardized, applying guidelines by Mohammed (2007) and using the Kurdish Sign dictionaries (Nashat Salim et al., 2013; Ghazi Dizayee, 2000) in the KRI education programs show some efforts towards KuSL standardization.

We develop a Kurdish Sign lexicon using the Kurdish Sign Language Dictionary (KuSLD) (Ghazi Dizayee, 2000), which is used in KRI.

Currently, no Kurdish Sign corpus is available, hence we aim at making Sorani texts sign-supported. That is, in the text conversion process we follow the spoken language and not the sign language structure. Sorani texts are mostly written in Persian-Arabic script (Hassani, 2018) hence we use the developed Kurdish Sign lexicon to make this type of the Sorani texts sign-supported.

The rest of this paper is organized as follows. Section 2. provides a brief background on sign language processing, Section 3. reviews the related work, Section 4. presents our approach, Section 5. illustrates the developed dataset, Section 6. discusses the results, finally, Section 7. concludes the paper.

2. Sign Language Processing

Sign languages are considered as genuine languages that place them among the minority languages (Senghas and Monaghan, 2002). Since sign languages consist

of visual gestures rather than voice as it is in spoken languages. The analysis and feature extraction of the former significantly differ from latter languages. However, for some languages, a variant of sign language also exists that follows the spoken/written language grammar, which is called sign-supported language (Elliott et al., 2008). The development of this variant is less challenging in the absence of required sign corpora and language models. The outcome could be used in various experimental and real-life occasions.

Several approaches exist to process sign languages. In the following sections, we discuss those approaches which are more related to our current stage of research.

2.1. Notation Systems

The sign visual gestures are normally denoted by special notations in order to be able to process them. Different notation systems are used to capture these gestures. The most popular ones are Stokoe, SignWriting, and HamNoSys.

Stokoe was one of the earliest attempts for a sign language notation system (McCarty, 2004). However, it was only concerned with manual sign representation, and it lacked any consideration for non-manual signs, such as eye gaze and shoulders movements, which are an essential entity to convey meaning by facial expression.

SignWriting represents the signed gestures spatially in a 2D canvas (Bouzid and Jemni, 2013b). It is designed to facilitate communication among the hearing-impaired community.

HamNoSys (Hamburg Notation System) is a phonetic translation system with iconicity, extensibility, and formal syntax characteristics used to denote sign languages (Hanke, 2004).

A comparative analysis by (Dhanjal and Singh, 2019) concluded that HamNoSys is the most widely used notation system for a variety of sign languages. HamNoSys symbols are available as a Unicode font (Hanke, 2004). This Unicode font symbolizes manual sign gestures and allows the generation of the signs by dividing the description into the handshapes, orientations, locations, and actions.

2.2. Markup Languages

To provide computer encoding for sign languages and to make their processing more efficient, several adoptions of the Extensive Markup Language (XML) have been suggested based on various sign notation systems.

The Sign Writing Markup Language (SWML) is a markup language proposed by da Rocha Costa and Dimuro (2001) based on SignWriting.

HamNoSys uses Signing Gesture Markup Language (SiGML), which gives a special XML tag to each HamNoSys symbol.

These markup languages are used in different applications, for instance, to be given to a 3D avatar to animate the signs.

3. Related Work

The only work on Kurdish Sign language processing that we were able to retrieve was by Hashim and Alizadeh (2018) wherein the researchers reported on their project on Kurdish Sign language recognition. That project focused on the recognition of Kurdish manual alphabets.

Therefore, as literature does not report on active studies on Kurdish Sign language processing, we review the topic in the context of other languages.

Sugandhi and Kaur (2018) introduced an online multilingual dictionary for avatar-based Indian Sign language. The system is designed to accept input from two languages English and Hindi. The input is transliterated into Hindi and then goes through the parser to be translated into Indian Sign Language (ISL). After extracting the root words of the input script, the target Hamburg Notations are retrieved from the database and converted into its corresponding SiGML. The generated SiGML is the input parameter for the Animation server, which uses Web Graphics Library (WebGL) for the avatar representation.

Aouiti (2013) proposed an approach to convert Arabic text into Arabic Sign language. The approach used an Arabic sentence/Sign language corpus as a core entity. The corpus includes Arabic sentences that were aligned with their corresponding sign representation. This helped to ensure that the represented sign refers to the real meaning of the input text. Afterward, the target sentence was syntactically and semantically analyzed by applying techniques, such as Morphological, Syntactic, Semantic, and Pragmatic analysis, which led to the generation of the glosses. The sign for each gloss was extracted from the corpus, which was sent to the avatar to be played.

Bouzid and Jemni (2013a) developed an avatar-based system to enhance the usability and readability of notation systems for deaf people. The system was developed using SignWriting (SW) notation and its markup language. Their focus was to make the path easier for hearing-impaired people to understand and represent signs in a written format. Since SW is presented in a 2D format and it is easy to guess the target gestures from the written notations, this helps hearing-impaired people to learn different sign languages depending on the SW notations. SW is designed for daily communication purposes rather than linguistic and corpus development and processing.

An automated reading system for SignWriting representation of Brazilian Sign language was introduced by Stiehl et al. (2015). They focused on SignWriting of several Brazilian signs and classified the symbols into several categories. Again, their purpose was to build a database of SignWriting representation for Brazilian Sign Language in order to involve hearing-impaired people into learning the notations and enable them to communicate with each other. This approach can also be used to have books, newspapers, dictionaries and such that are written in notation symbols and can be understood by hearing-impaired people or sign learn-

ers.

To summarize, we follow the approach of Sugandhi and Kaur (2018) because of two reasons. First, because HamNoSys is a proper method for corpus development, and second because SignWiriting is majorly used for the communication among the hearing-impaired community and not between the hearing-impaired community and people with no hearing difficulties.

4. Method

We develop a dataset based on KuSLD. We also prepare and adapt a tool to translate Sorani texts into Kurdish Sign language to be animated by an avatar.

We prepare the HamNoSys notations manually by analyzing the gestures available in KuSLD. We use the Ham2HPSG and eSIGN (Hanke and Popescu, 2003) to create the dataset and extract the SiGML codes.

We implement a sign-supported tool based on the architecture that is shown in Figure 1. In this architecture, the Language Model (LM), in its current form, is the developed dataset, which could be considered as the Kurdish (Sorani) sign lexicon. The input text goes through the tokenization process to extract the meaningful components from it.

Similar to the existing sign-supported tools for other languages, the translation is word-by-word for the words that are found in the LM. Otherwise, the word will be replaced by a sequence of its letters in the sign language. Then the whole text is compiled into SiGML files, which will be sent to an avatar to be animated.

We evaluate the tool by feeding it with input of four categories, namely alphabets, numbers, words, and sentences. The tool then plays the translation to the human individuals who are either hearing-impaired or Kurdish Sign language educators. As subjective understanding is not accurate (Kipp et al., 2011), we ask the testers to write down their understanding. We calculate the accuracy by the percentage of correctly understood cases for played alphabets, numbers, words, and sentences.

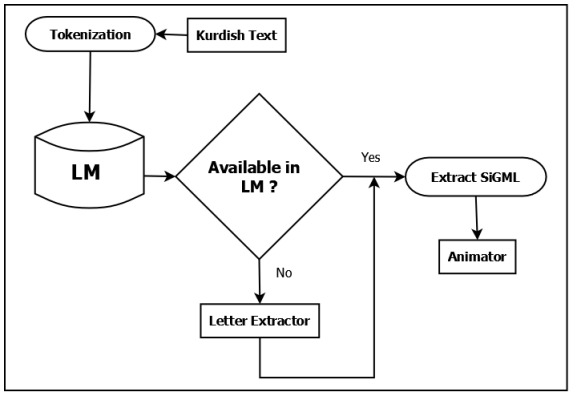


Figure 1: Kurdish text to sign proposed architecture

5. Developed Dataset

The KuSLD consists of 2315 different sign gestures from 38 different categories. Our dataset, currently, consists of 20% of each category. However, we converted the alphabet and numbers completely. This adds more entries to the dataset, which sums up to approximately 560 entries.

The KuSLD categories are listed in Table 1. A sample for the prepared HamNoSys for Kurdish letters and words is shown in Figures 2 and 3.

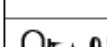
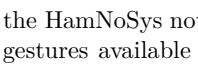
HamNoSys	Word
	من/Me
	سڵاو/Hello
	ئەمڕۆ/Today
	کتێب/Book
	کۆمپیوتەر/Computer
	خوێندن/Read

Figure 2: HamNoSys Sample for Kurdish words

HamNoSys	Letter
	ئ/E
	ا/A
	ب/B
	پ/P
	ت/T
	ج/C

Figure 3: HamNoSys Sample for Kurdish letters

We extracted the generated SiGML for the corresponding HamNoSys dataset, which was sent to an avatar to be animated. Two samples of extracted SiGML for the letter "ب" (B) and the word "زانکۆ" (University) in Kurdish are shown in Figures 4 and 5.

6. Findings and Discussion

We played a sample of letters of the prepared dataset to the hearing-impaired individuals. The test showed a 100% understanding of the test data.

The results of playing words showed a 65% correct understanding of the played words.

The accuracy of the tool for understanding sentences was approximately 30%.

In the evaluation process, the person could recognize all shown letters successfully since they are clearly

Category	Entries
Kurdish Alphabets and Grammar	96
Greetings	13
Asking Questions	20
Colors	14
Time and Days	45
Opposites	134
Biology and Animals	105
Fruits and Vegetables	54
Food and drinks	87
Home	148
Family and Relatives	34
Human Body	48
Health and Hygiene	78
Clothes and Cosmetics	70
People with Special Needs	18
School	60
Verbs	193
Engineering and Mathematics	95
Science and Measuring Quantities	65
Art and Literature	25
Sound and Music	23
Sports and Games	56
History	27
Geography	50
Countries	78
City and Places	132
Transportation	63
Agriculture	46
Jobs	69
Economy	16
Finance and Currency	19
General Services	58
Institutes	17
Religion and Believes	52
Politics	41
Famous Personalities	11
Special Events	31
Other Words	124
Total	2315

Table 1: Kurdish Sign Language Dictionary Categories

shown in the dictionary. On the other hand, the signs for the words had a lower evaluation outcome. The person could not understand some of the words.

One reason for this was the usage of two different sign dictionaries in KRI. One of these dictionaries represents all signs based on the lexicon description, while the other (Ghazi Dizayee, 2000) uses vocal description for some of its entries. Our dataset was developed based on the latter. Both dictionaries are used interchangeably, but they provide different representations for specific signs depending on the context where they appear. This issue also affected sentence evaluation. Also, since we used a word by word translation, the hearing-impaired person was unable to understand

```
<sigml>
        <hns_sign gloss=<"ب ">
                <hamnosys_manual>
                        <hamflathand/>
                        <hamextfingeril/>
                        <hampalmdr/>
                        <hamtongue/>
                </hamnosys_manual>
 </hns_sign>
</sigml>
```

Figure 4: SiGML sample for letter "ب"

```
<sigml>
        <hns_sign gloss= <"زانکۆ">
                <hamnosys_manual>
                        <hamcee12/>
                        <hamthumbopenmod/>
                        <hamextfingerul/>
                        <hampalmd/>
                        <hamchin/>
                        <hamparbegin/>
                        <hamreplace/>
                        <hamextfingerul/>
                        <hampalmu/>
                        <hamparend/>
                        <hamlrbeside/>
                        <hamcheek/>
                </hamnosys_manual>
 </hns_sign>
</sigml>
```

Figure 5: SiGML sample for word "زانکۆ"

the meaning of a majority of the sentences as a whole. Therefore, the sentence evaluation achieved low accuracy, which is typical for the sign-supported systems.

7. Conclusion

We used HamNoSys to develop a sign dataset and its equivalent SiGML for Kurdish. We chose HamNoSys over SignWriting because of our plan to develop Kurdish Sign corpora in the future.

Our developed dataset includes approximately 560 entries consisting of the alphabet, numbers, and words. We also implemented a tool to translate Sorani texts into the Kurdish Sign language, which could be animated by an avatar.

We evaluated the tool by showing the animated output to hearing-impaired persons on the three aspects of understanding the sign gestures, namely letters, words, and sentences. The test showed a 100% understanding for the letters, a 65% for isolated words, and approxi-

mately 30% for sentences.
The main reasons for the low accuracy were the usage of more than one sign dictionary in the target community and the word-by-word translation of the input texts.
As future work, we are targeting the development of a language model based on the grammar of the Kurdish Sign language. Additionally, we aim to add more entries to the developed dataset. Furthermore, we would like to include other Kurdish dialects in the dataset.

8. Acknowledgements

We would like to appreciate the assistance of the Erbil Hearing-impairment Association for their assistance to access the Kurdish Sign dictionaries. Also, we acknowledge the help of Hiwa Center for Deaf and Mute in Erbil, particularly Ms. Mehan Fatah and Ms. Shno Aziz, for providing us with Kurdish Sign language resources and for their help in the evaluation process. Furthermore, we appreciate the constructive feedback we received from anonymous reviewers, which has helped us to improve the quality of the paper.

9. Bibliographical References

Ahmadi, S., Hassani, H., and McCrae, J. P. (2019). Towards Electronic Lexicography for the Kurdish language. In *Proceedings of the sixth biennial conference on electronic lexicography (eLex)*. eLex 2019.

Aouiti, N. (2013). Towards an automatic translation from arabic text to sign language. In *Fourth International Conference on Information and Communication Technology and Accessibility (ICTA)*, pages 1–4. IEEE.

Bouzid, Y. and Jemni, M. (2013a). An animated avatar to interpret SignWriting transcription. In *2013 International Conference on Electrical Engineering and Software Applications*, pages 1–5. IEEE.

Bouzid, Y. and Jemni, M. (2013b). An Avatar based approach for automatically interpreting a sign language notation. In *2013 IEEE 13th International Conference on Advanced Learning Technologies*, pages 92–94. IEEE.

Cooper, H., Holt, B., and Bowden, R. (2011). Sign language recognition. In *Visual Analysis of Humans*, pages 539–562. Springer.

Cormier, K., Fox, N., Woll, B., Zisserman, A., Camgöz, N. C., and Bowden, R. (2019). ExTOL: Automatic recognition of British Sign Language using the BSL Corpus. In *Proceedings of 6th Workshop on Sign Language Translation and Avatar Technology (SLTAT) 2019*. Universitat Hamburg.

da Rocha Costa, A. C. and Dimuro, G. P. (2001). SignWriting-Based Sign Language Processing. In *International Gesture Workshop*, pages 202–205. Springer.

Dhanjal, A. S. and Singh, W. (2019). Comparative Analysis of Sign Language Notation Systems for Indian Sign Language. In *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)*, pages 1–6. IEEE.

Elliott, R., Glauert, J. R., Kennaway, J., Marshall, I., and Safar, E. (2008). Linguistic modelling and language-processing technologies for Avatar-based sign language presentation. *Universal Access in the Information Society*, 6(4):375–391.

Halawani, S. M. (2008). Arabic sign language translation system on mobile devices. *IJCSNS International Journal of Computer Science and Network Security*, 8(1):251–256.

Hanke, T. and Popescu, H. (2003). Intelligent sign editor. eSIGN project deliverable. *D2*, 3:2003.

Hanke, T. (2004). HamNoSys-representing sign language data in language resources and language processing contexts. In *LREC*, volume 4, pages 1–6.

Hashim, A. D. and Alizadeh, F. (2018). Kurdish Sign Language Recognition System. *UKH Journal of Science and Engineering*, 2(1):1–6.

Hassani, H. and Medjedovic, D. (2016). Automatic Kurdish dialects identification. *Computer Science & Information Technology*, 6(2):61–78.

Hassani, H. (2018). BLARK for multi-dialect languages: towards the Kurdish BLARK. *Language Resources and Evaluation*, 52(2):625–644.

Jepsen, J. B., De Clerck, G., Lutalo-Kiingi, S., and McGregor, W. B. (2015). *Sign languages of the world: A comparative handbook*. De Gruyter.

Kelly, D., Reilly Delannoy, J., Mc Donald, J., and Markham, C. (2009). A framework for continuous multimodal sign language recognition. In *Proceedings of the 2009 international conference on Multimodal interfaces*, pages 351–358.

Kipp, M., Heloir, A., and Nguyen, Q. (2011). Sign language avatars: Animation and comprehensibility. In *International Workshop on Intelligent Virtual Agents*, pages 113–126. Springer.

Liwicki, S. and Everingham, M. (2009). Automatic recognition of fingerspelled words in British Sign Language. In *2009 IEEE computer society conference on computer vision and pattern recognition workshops*, pages 50–57. IEEE.

McCarty, A. L. (2004). Notation systems for reading and writing sign language. *The Analysis of verbal behavior*, 20(1):129–134.

Mohammed, M. F. (2007). *The Sign Language for Deaf [In Arabic]*. Ministry of Labor and Social Affairs, Iraq.

Senghas, R. J. and Monaghan, L. (2002). Signs of their times: Deaf communities and the culture of language. *Annual Review of Anthropology*, 31(1):69–97.

Stiehl, D., Addams, L., Oliveira, L. S., Guimarães, C., and Britto, A. (2015). Towards a SignWriting recognition system. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 26–30. IEEE.

Sugandhi, P. K. and Kaur, S. (2018). Online Multilinguial Dictionary Using Hamburg Notaiton for Avatar-Based Indian Sign Language Generation System. *Int. J. Cogn. Lang. Sci.*, 12(8):1116–1122.

Wazalwar, S. S. and Shrawankar, U. (2017). Interpretation of sign language into English using NLP techniques. *Journal of Information and Optimization Sciences*, 38(6):895–910.

10. Language Resource References

Ghazi Dizayee, Adwiya. (2000). *Sign Dictionary for Hearing-Impairment [In Kurdish]*. The independent Human Rights Commission- Ministry of Labor and Social Affairs of Kurdistan Regional Government - Iraq- UNICEF- MEDS Organisation.

Nashat Salim, Nazanin and Hama Rashid, Jamil and Haji Omar, Salam. (2013). *Sign Dictionary for Hearing-Impairment [In Kurdish]*. The independent Human Rights Commission- Ministry of Labor and Social Affairs of Kurdistan Regional Government - Iraq.

Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 123–126
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

Recognition of Static Features in Sign Language Using Key-Points

Ioannis Koulierakis[1], Georgios Siolas[1], Eleni Efthimiou[2], Stavroula-Evita Fotinea[2], Andreas-Georgios Stafylopatis[1]

[1]National Technical University of Athens, School of Electrical and Computer Engineering,

Intelligent Systems Laboratory,

[2]Sign Language Technologies Team, Department of Embodied Interaction and Robotics,

Institute for Language and Speech Processing (ILSP) / ATHENA RC

Zografou Campus, 9, Iroon Polytechniou str, 15780 Zografou, Greece,

Artemidos 6 & Epidavrou, 15125 Maroussi, Greece

koulyia@gmail.com, gsiolas@islab.ntua.gr, {eleni_e, evita}@athenarc.gr, andreas@cs.ntua.gr

Abstract

In this paper we report on a research effort focusing on recognition of static features of sign formation in single sign videos. Three sequential models have been developed for handshape, palm orientation and location of sign formation respectively, which make use of key-points extracted via OpenPose software. The models have been applied to a Danish and a Greek Sign Language dataset, providing results around 96%. Moreover, during the reported research, a method has been developed for identifying the time-frame of real signing in the video, which allows to ignore transition frames during sign recognition processing.

Keywords: Sign language recognition algorithm, deep neural networks, key-point extraction

1. Introduction

One of the problems relating to sign language recognition is the lack of appropriate datasets for algorithm training, since most datasets are recorded for academic purposes and as such, they concentrate in human learning rather than machine learning. Therefore, most data collections contain a very large number of different glosses with very few repetitions of each. This characteristic makes it very unlikely for these datasets to be used as training sets for classification algorithms in sign recognition level. Thus, we developed a system in the direction of "phonological" features recognition. This way we can extract a dataset with a lot of examples for every handshape, palm orientation and hand location out of the video collections.

2. Datasets

For the purposes of the project two collections of single gloss videos were used as datasets.

The first one is "Noema +" which was developed by the Greek Institute for Language and Speech Processing (ILSP), Athena. It contains approximately 3000 lemmas of the Greek Language were signed by one native Greek signer and many of them are recorded two or three times. The total amount of videos is 3195 annotated with HamNoSys (Hanke, 2004).

The second one is the "Danish Sign Language Dictionary". It was developed and edited at the Centre for Sign Language and Sign Supported Communication – KC in close cooperation with the Danish Deaf Association (DDL) Centre for Sign Language as a dictionary of the Danish Sign Language (DSL). The dictionary is consisted by single-sign videos as well as videos including short sentences in DSL. We used the single gloss videos which are 2714 in total, signed by several different signers. All these videos are annotated with a variation of HamNoSys that uses only one descriptor per instance (handshape, location, etc). For the description of the handshapes, 69 different names were used. Most of them are named after a letter of the Danish fingerspelling alphabet. Based on them all videos were annotated. This feature is making the whole process easier when trying to split the dictionary into handshape classes.

3. Openpose

OpenPose is a software freely distributed by Carnegie Melon University, Perceptual Computing Lab (Cao et al., 2018). It is used as a tool of human body keypoints extraction from a single image or video frame. It offers an estimation of 25 body/foot keypoints, 2x21 hand keypoints and 70 face keypoints. In the case of a 2D video input, for each keypoint it returns a vector containing 3 elements. The first 2 correspond to the (x,y) coordinates with reference to the upper left corner of the image. The third is a value in the range [0,1] which is quantification of the confidence given by the

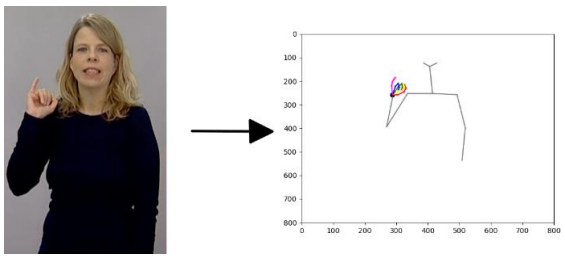


Figure 1: Example of OpenPose

program that the specific keypoint is correctly located in the frame. The novelty behind OpenPose relies on the fact that it works for more than one person per image but more importantly the keypoint analysis is not affected when part of the individual's body is out of frame. This last feature is crucial for applications on sign language videos where the signer appears above the waist level (Figure 1).

4. Our Method

The first step in our method is transforming each video frame into keypoints using the OpenPose software. With this step we keep all the necessary information of the signer's posture and hand articulation, while we reduce the data dimensionality from 1280 x 720 pixels to 137 keypoints. Moreover, in our case the keypoints of the legs are redundant since none of the videos shows the bottom half of the signer. In addition, for each of the systems we will analyze below, we used a different number of keypoints related to the feature we are trying to classify in each case.

In general, the complete feature vector produced by OpenPose has the form:

$$X_\tau = \begin{bmatrix} x_{0,\tau}\ y_{0,\tau}\ \sigma_{0,\tau} \\ x_{1,\tau}\ y_{1,\tau}\ \sigma_{1,\tau} \\ \vdots \\ x_{14,\tau}\ y_{14,\tau}\ \sigma_{14,\tau} \\ x^H_{0,\tau}\ y^H_{0,\tau}\ \sigma^H_{0,\tau} \\ x^H_{1,\tau}\ y^H_{1,\tau}\ \sigma^H_{1,\tau} \\ \vdots \\ x^H_{20,\tau}\ y^H_{20,\tau}\ \sigma^H_{20,\tau} \\ x^h_{0,\tau}\ y^h_{0,\tau}\ \sigma^h_{0,\tau} \\ x^h_{1,\tau}\ y^h_{1,\tau}\ \sigma^h_{1,\tau} \\ \vdots \\ x^h_{20,\tau}\ y^h_{20,\tau}\ \sigma^h_{20,\tau} \end{bmatrix}$$

Where $[x_{i,\tau}, y_{i,\tau}, \sigma_{i,\tau}]$ is the i^{th} keypoint of the τ^{th} frame of the video. With the superscripts H, h we denote the keypoints of the dominant and non-dominant hand, respectively.

4.1. Segmentation

The problem of training a model on our data is a problem of semi-supervised learning. The reason is that in every video the annotation provides us with information on which are the static phonological features appearing in the video and the order in which they appear, but we lack a matching of the static features with individual frames. Moreover, we need a filtering of transitional frames that represent none of the annotated features. Those frames appear when a signer starts or stops signing moving his/her hands from or to resting pose, or during transition from sign to sign or from handshape to handshape into one sign. In all those cases the frames have no use in our training algorithm. In (Koller et al., 2016) this problem is solved by considering a "junk" state for those frames and using an Expectation Maximization algorithm for finding the most probable alignment between the frames and the annotation. On the other hand, we will use an alternative method to what was proposed by (Ko et al., 2018). This method is relying on the work of (Choudhury et al., 2017) that categorizes the movement during signing in "Movement Epenthesis" and "Signing" based on the velocity of the centroid of the contour produced during the hand tracking stage. This method sets a velocity threshold and rejects every sequence of frames with greater velocity than the threshold.

In our method we are transforming each video frame into keypoints using the OpenPose software.
This help us skip the hand recognition and tracking process while maintaining the maximal accuracy provided by OpenPose. In addition, we have the opportunity to calculate the velocity of the hand based on more than one point of interest. We calculate the total hand velocity as the sum of the velocities of each keypoint between two frames on the basis of the following equation:

$$v_\tau = \sum_{i=0}^{20} \sqrt{\left(x^H_{i,\tau} - x^H_{i,\tau+1}\right)^2 + \left(y^H_{i,\tau} - y^H_{i,\tau+1}\right)^2}$$

We present our methodology based on the handshape recognition. Although, the method is outright extendable to the other two characteristics.
We modify the method for rejecting redundant frames based on velocity threshold and extend it by adding one more rule. Every frame is removed from dataset unless it satisfies the following 3 rules:

- Belong in a sequence of 5 frames with total velocity below a threshold T_v.
- The logarithmic sum of certainty σ_i of all points is over a threshold T_σ.
- Wrist is over the waist level (not a hand resting posture)

Both video datasets were recorded at 25 frames/second and so the 5 frame sequence corresponds to 0.2 seconds.
We remove the third column from the feature vector X_τ and we use the provided information in the second rule in order to remove bad quality data from our dataset before training. The third rule was added to remove frames from the start and end of each video where the signer is crossing his/her hands on waist level. These frames do not involve signing but they pass the first two rules due to very low movement and clarity.
Our final step is to match each frame remaining to the matching handshape. The advantage here is that the maximum number of different handshapes appearing in every video is 2 due to the fact that we have single gloss videos. If according to the HamNoSys annotation, only one handshape appears in the video, we know that the frames are representing that specific handshape. Otherwise, when two handshapes appear in the video, we are clustering the frames into two clusters using Gaussian Mixture Model (GMM) (Koller et al., 2016) (Theodorakis et al., 2014) (Pitsikalis et al. 2011). The first handshape is matched to the cluster the elements of which appear earlier by mean in the video and the second is matched to the other one.

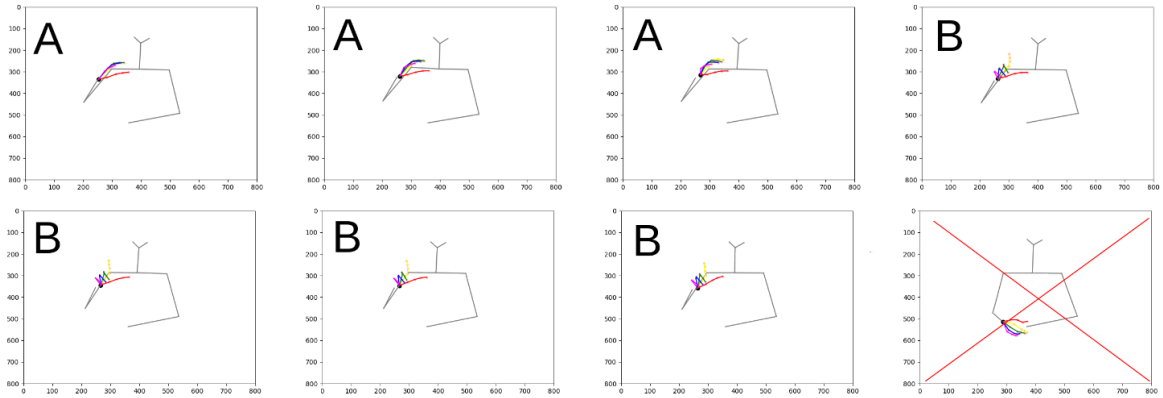


Figure 2: Example of Segmentation process

For example, in Figure 2, there is an example of 8 frames extracted from a video and transformed into keypoints. Beforehand we know that there are two different handshapes. During the segmentation, the last frame is rejected because the signer has crossed his hands and so the right palm is below the elbow level. Otherwise this frame would be labeled as or as adding a false element into the dataset. The GMM algorithm splits the frames into clusters A,B. Cluster-A appears earlier in the sequence so we label the elements of cluster-A, cluster-B as , , respectively.

4.2. Training

At this point we have created a collection of frames representing each possible handshape. In our final dataset, 66 different handshapes, 12 different palm orientations, 33 different locations have found after the segmentations. Approximately 60000 frames ended in the final segmentation for each of the three static features. For the handshape training we used the keypoints extracted only from the dominant hand. We changed the reference system fixing the centroid of the 21 keypoints to (0,0) and the feature vector was normalised using the mean and standard deviation for each dimension, respectively. Moreover, the same process was used for creating the palm-orientation and location feature vectors. For the palm orientation we used all the upper body keypoints plus the dominant hand. Finally, for the location feature vector we included every keypoint including the non-dominant. The reason for this extension of the feature vectors was due to the fact that the orientation of the body of the signer is not the same for all videos and the hand orientation has to be recognized relatively to the body. Obviously, all keypoints are necessary for palm location due to the fact that every articulation is described relatively to a body part including the non-dominant hand and the face. For the two models we use Multi-Layer Perceptrons (MLPs) since we classify each frame independently. 5 hidden layers models with 128 neurons in each layer and softmax activation function are used both for handshape and orientation classifiers. In total we train our models for approximately 400 epochs.

For the training we isolated the 25% of all the videos as a test set. In this way the models are tested in classifying features from signs they have not come up with or even signers for DSL dataset that the systems are not trained on.

4.3. Results

	Train Set	Val. Set	Test Set
Handshape	99.8%	95.7%	95.6%
Palm Orientation	99.7%	96.2%	96.1%
Location	99.8%	97.1%	96.8%

Table 1: Model Accuracies

In Table 1 we can see that the final accuracies of every model is over 95%. According to Figures 3,4,5 the models can almost perfectly classify the training set. We should, also, point out that many of the errors in our classification method could be related to errors during the segmentation.

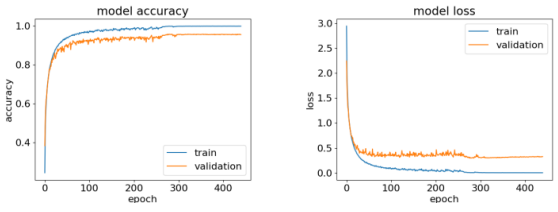


Figure 3: Handshape recognition model

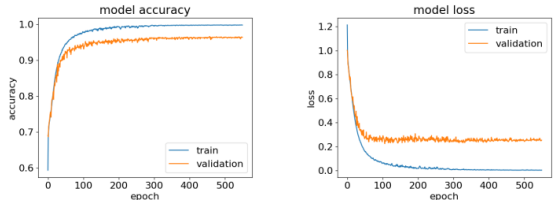


Figure 4: Palm orientation recognition model

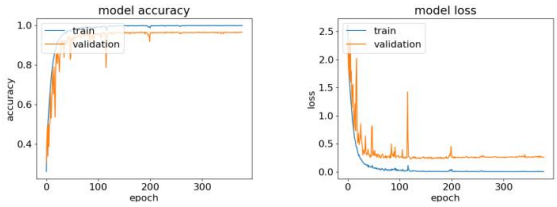


Figure 5: Hand Location recognition model

5. Conclusion

Research efforts relating to recognition of static features of sign formation including the handshape, the palm orientation and the location of signs by means of sequential models, have provided encouraging results as shown in 4.3 above. Such results may prove especially helpful towards (semi-)automatic annotation of SL videos. Furthermore, embedding of the three models handling handshape, palm orientation and location of sign in recurrent neural networks is expected to pave the way towards continuous SL recognition.

6. Acknowledgements

The authors acknowledge support of this work by the project “Computational Science and Technologies: Data, Content and Interaction” (MIS 5002437), which is implemented under the Action “Reinforcement of the Research and Innovation Infrastructure”, funded by the Operational Programme "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund). Moreover, they wish to thank Center for Tegnsprog (2008–2018). Ordbog over Dansk Tegnsprog. http://www.tegnsprog.dk for providing the Danish Sign Language dataset used in a number of experiments in the framework of the here presented research.

7. References

Ananya Choudhury, Anjan Kumar Talukdar, Manas Kamal Bhuyan, and Kandarpa Kumar Sarma (2017). Movement Epenthesis Detection for Continuous Sign Language Recognition. Journal of Intelligent Systems.

Oscar Koller, Hermann Ney and Richard Bowden, (2016). Automatic Alignment of HamNoSys Subunits for Continuous Sign Language Recognition.

Sang-Ki Ko, Chang Kim, Hyedong Jung, and Choongsang Cho (2019). Neural sign language translation based on human keypoint estimation. Applied Sciences.

Stavros Theodorakis, Vassilis Pitsikalis, and Petros Maragos (2014). Dynamic-static unsupervised sequentiality, statistical subunits and lexicon for sign language recognition. Image and Vision Computing.

Thomas Hanke (2004). HamNoSys - Representing Sign Language Data in Language Resources and Language Processing Contexts. Lrec 2004.

Vassilis Pitsikalis, Stavros Theodorakis, Christian Vogler, and Petros Maragos (2011). Advances in phnetics-based sub-unit modeling for transcrition alignment and sign language recognition.

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-EnWei, and Yaser Sheikh (2018). OpenPose: realtime multi-person 2D pose estimation usng Part Affinity Fields.

Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 127–134
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

Collocations in Sign Language Lexicography: Towards Semantic Abstractions for Word Sense Discrimination

Gabriele Langer, Marc Schulder
Institute for German Sign Language
University of Hamburg, Germany
gabriele.langer@uni-hamburg.de, marc.schulder@uni-hamburg.de

Abstract

In general monolingual lexicography a corpus-based approach to word sense discrimination (WSD) is the current standard. Automatically generated lexical profiles such as Word Sketches provide an overview on typical uses in the form of collocate lists grouped by their part of speech categories and their syntactic dependency relations to the base item. Collocates are sorted by their typicality according to frequency-based rankings. With the advancement of sign language (SL) corpora, SL lexicography can finally be based on actual language use as reflected in corpus data. In order to use such data effectively and gain new insights on sign usage, automatically generated collocation profiles need to be developed under the special conditions and circumstances of the SL data available. One of these conditions is that many of the prerequesites for the automatic syntactic parsing of corpora are not yet available for SL. In this article we describe a collocation summary generated from *DGS Corpus* data which is used for WSD as well as in entry-writing. The summary works based on the glosses used for lemmatisation. In addition, we explore how other resources can be utilised to add an additional layer of semantic grouping to the collocation analysis. For this experimental approach we use glosses, concepts, and wordnet supersenses.

Keywords: collocations, sign language lexicography, corpus-based analysis of sign usage, lexical profile, word sense discrimination, sign language NLP, cross-lingual bootstrapping

1. Introduction

One central task in the lexicographic analysis of a language is to identify different word senses, i. e. different uses of a given word or expression.[1] In corpus-based lexicography, this is achieved by inspecting many occurrences of the word of interest in their linguistic context to determine their contextual meanings and conditions of use. Occurrences of same or similar meanings and usages are then grouped together as representing a specific use of the expression type. Each use that is identified in this manner is then described as a particular sense in the dictionary entry. It is necessary to consider as many examples of usage as possible in order to identify and substantiate the existence of specific word senses and prevent overlooking typical uses.

During lexicographic analysis, single occurrences and their immediate linguistic context are usually reviewed in a concordance view of the data (also known as *Keyword in Context* or *KWIC*). However, for high frequency words in large corpora it is impossible to inspect all existing occurrences manually. Instead, lexicographers either resort to inspecting a randomised sample of occurrences from different sources (Landau, 2001, p. 296) or to consulting the output of a lexical profiling software. Lexical profiles such as the Word Sketches generated by Sketch Engine are summaries of collocate lists grouped by part of speech (POS) tag sequences. As "[...] the regular lexical environment of a word is one of the most reliable indicators if its senses", they provide lexicographers with a "[...] preferred starting point to their analyses of complex headwords." (Atkins and Rundell, 2008, p. 110-111).

In sign language linguistics, corpus-based lexicography is still a very new field. In general, sign languages (SL) belong to the less researched and less resourced languages and are without a written form and tradition. This presents technical and methodological challenges when collecting, annotating and analysing SL texts in a corpus and when describing signs and their uses in dictionaries (Zwitserlood et al., 2013). For instance they lack sufficient numbers of sources to sample from, as well as most natural language processing (NLP) tools, such as automatic POS taggers.

To enable corpus-based lexicographic work and research on usage in SL it would be very useful to have concordance views and lexical profiles with collocational information of the target sign. Without a direct written representation of signing, ways of representing, grouping and ranking occurrences must be found that do not rely on an unmediated written representation of the language samples under investigation. In SL corpus linguistics the most feasible way is to use the gloss annotation for all aspects of searching, sorting and counting while time-aligned video files stay easily accessible for viewing the actual language samples in their direct, unmediated form (cf. Johnston, 2010).

In this article we explore such an approach. Section 2 provides relevant background on sign language linguistics, while Section 3 introduces the corpus data used in our work. Section 4 gives general information on collocations. In Section 5 we introduce a collocation view based on sign glosses, while in Section 6 we explore the use of automatic cross-lingual methods for providing an additional semantic collocation layer, based on wordnet supersenses. Section 7 provides an outlook on possible future extensions of our work.

[1] For the purpose of brevity, when talking about abstractions that refer to both signed and spoken languages we use terminology from lexicography and linguistics which in some cases reflects the spoken language bias of its origin. When referring to the general concept of a lexical unit, the term *word* covers both the word of a spoken language and the sign of a signed language. Similarly, *word sense* also covers *sign sense* and *phonological variant* covers *cherological variant*.

2. Background

In this section we provide a brief overview of sign language research, specifically regarding language transcription (Section 2.1), the state of research on general linguistic concepts (Section 2.2) and the availability of natural language processing software (Section 2.3).

2.1. Transcription of Sign Languages

Since sign languages lack a standard form of written representation and the usual writing systems are not suitable to represent signing, it is a common method in SL research to represent signing in a stable, non-perishing form via the use of glosses. In SL corpus research a gloss is a metalinguistic label to represent a specific sign type. Each sign type is assigned a unique gloss that is used consistently and exclusively for that sign (Johnston, 2010). A gloss is comprised of a spoken language word (the *gloss name*, usually written in capital letters) and additional suffixes such as digits and letters to distinguish between different signs and variants.

The purpose of the gloss as a label is that it can be written, read and easily remembered by humans as well as sorted, counted and manipulated by computers.[2] Through the representation of sign texts as glosses certain NLP operations can be performed on these representations. However, glosses by themselves do not reveal anything about the actual sign forms. For this they need to be related and (preferably) directly linked to a corresponding lexical entry. Furthermore, glosses are not to be mistaken as context-appropriate translations of a sign. Using a spoken language word as a label for a sign bears the risk of unsuitably inferring that syntactic and semantic properties of the spoken language word, such as parts of speech and nuances of meaning, also are valid for the sign of the sign language (cf. Slobin, 2008).

Bearing these drawbacks and risks in mind, gloss names can nevertheless be used for NLP purposes as a general, rough and incomplete approximation to a sign's general meaning, as gloss names are usually chosen so that the spoken language word indicates a core meaning of the respective signs (see, for example, Johnston (2010, pp. 119–120)).

2.2. The State of Sign Language Research

One of the major challenges of linguistic research for sign languages is that there still are no commonly agreed upon theories for certain basic categories and concepts, such as sentences and parts of speech. As languages without a written form and primarily used in face-to-face communication, sign languages share some structural properties with spoken language mode (as opposed to written language), for example the difficult issue of determining sentence boundaries. While written language is usually segmented based on orthographic markers, the spoken language mode of spoken languages has been very difficult to segment (Auer, 2010; Westpfahl and Gorisch, 2018) and thus segmentation in many speech corpora is based solely on pauses in speech (Hamaker et al., 1998; Schmidt, 2014). In addition, the visual modality of SL informs very different language structures that still lack comprehensive analysis and description. As Schwager and Zeshan (2008) observe, "[b]y and large, it has not been easy to identify workable syntactic tests for sign languages, given that they often have relatively free word order and some of their sentence structures are unfamiliar from a spoken language background, including spatial syntax and simultaneous constructions." For instance, in German Sign Language many signs can be multi-functional, appearing for example as either predicate or argument. The body of research does not yet provide a commonly agreed background on sentence structures and POS categories that can robustly be used for segmenting sentences and tag signs for POS in corpus annotation.

2.3. NLP for Sign Languages

Presently, sign language research has to manage without most of the NLP tools and procedures that are ubiquitous for well researched and resourced spoken languages like English or German. Apart from the general struggles of any less resourced language, such as a lack of (machine-readable) language data, sign language research is made especially challenging by its specific modality in the visual-gestural domain, the resulting very different language structures, the lack of comprehensive grammars that categorise and describe these structures and could be used for widespread analyses, and the absence of large machine-readable lexical resources.

One reason for the lack of NLP tools for sign languages is that for almost any NLP task, a series of other tasks must be executed in preparation for it. The most common of these pre-processing steps are a) sentence tokenisation (splitting a text into sentences), b) word tokenisation (splitting a sentence into words or signs), c) part of speech tagging, and d) lemmatisation (turning a word or sign into its citation form). Even these steps build upon one another (e. g. the lemma of a word depends on its part of speech). As we discussed in Section 2.2, neither sentence boundaries nor parts of speech are even fully defined for sign languages, so designing machine classifiers for them (let alone for more advanced tasks) is not yet feasible. NLP for sign language research therefore mostly consists to makeshift solutions, broad generalisations and solutions bootstrapped from spoken language data and tools.

3. Data for German Sign Language

3.1. The DGS Corpus

For lexicographic analyses and descriptions, such as the collocation analyses discussed in this article, we exclusively use the data of the *DGS Corpus*. This corpus of filmed natural and near-natural conversations and narrations signed in German Sign Language (DGS) has been collected in Germany between 2010 and 2012 (Nishio et al., 2010). It includes language samples of 330 persons from all over Germany who were filmed in pairs. 560 hours of signing were recorded, of which about 79 hrs of running text have been annotated and lemmatised in iLex so far. As of February 2020, the corpus contains close to 530,000 token tags. About 50 hours of the material have been published as the *Public DGS Corpus* and are available online

[2] In the case of iLex, instead of using unique glosses the iLex-internal unique ids can also be used for identification, counting and other computational tasks. However, in the following we describe the approach as based on the glosses.

through the community portal *MY DGS*[3] and the research portal *MY DGS – annotated*[4]. Most lemmatised running texts in the *DGS Corpus* data are translated into German and time-aligned with utterances as suggested by the German translation. For collocation analyses and lexicographic description all lemmatised corpus data is used, including lemmatised data of the unpublished material.

3.2. The iLex Database of the IDGS

The work described in this article is conducted in the environment of iLex, an integrated annotation tool and lexical database (Hanke and Storz, 2008). During lemmatisation tokens are matched to types that are defined as lexical entries. A token can be lemmatised only after the respective type has first been established as a lexical entry. Thus, annotation in iLex consequently leads to the identification of sign types and their description in lexical entries.

At the Institute of German Sign Language and Communication of the Deaf (IDGS) at Hamburg University we use a complex structure of **types** and subordinate **subtypes** for token-type matching during lemmatisation called *double glossing* (Langer et al., 2016; Konrad et al., 2018).[5]

A type is an abstract unit of the sign language with a specific form often associated with an underlying image – that can have several differing realisations in actual use – and with one or several conventional meanings. Each subtype belongs to a specific type and roughly represents one of its conventional uses. The conventional meaning that a subtype covers is described by linking one or more **concepts** to the particular subtype entry. *Concepts* here are to be understood as pre-theoretical and pre-analytical indications of conventional meanings. In iLex they are their own entities that are identified by a German word or expression covering the conventional meaning of the sign. When the German word is ambiguous, a disambiguating **context** can be added to specify the meaning.

Often the words that are chosen as concept descriptions to indicate a conventional meaning of the subtype correspond to a specific mouthing associated with the sign. For signs that are presumed to be multifunctional, several concepts for the same German root may be specified to represent their POS-specific forms (e. g. *"Arbeit"* (noun) and *"arbeiten"* (verb)).

During lemmatisation sign tokens are matched to a type or one of its subtypes. Each type and subtype is identified by a internal unique id, a citation form described in HamNoSys notation (Hanke, 2004), and – for the benefit of the human user – receives a unique gloss. iLex ensures that glosses for types and subtypes are unique.

The iLex database used and maintained by the IDGS comprises data from several projects. Data from all included projects commonly use the same type and subtype entries for lemmatisation. The DGS-Korpus project (cf. Section 3.1) re-uses the sign type entries established in previous projects with their descriptions as a starting point and adds, re-evaluates and corrects them when necessary.

We exclusively use tokens from the *DGS Corpus* data for collocational analyses, but the sign type entry information – namely gloss names and concepts – may stem from previous projects. As the linking of concepts to subtypes was done primarily on the basis of introspection and in accordance with the different guidelines of these projects, this information can be of varying quality.

The basic annotation of the *DGS Corpus* does not attempt to identify and tag segments of DGS other than individual signs. The only available approximation to DGS sentences or utterances are the German translation tags. German translations are source-language oriented translations, but they still constitute sentences or utterances in German. As part of the annotation they are time-aligned to the DGS signing in the video and token tags in the transcript. Segment lengths were determined pragmatically while as many of the DGS structures and boundary signals as possible were taken into consideration (cf. Section *Translation into German* in Konrad et al. (2018)). Needless to say that this is not an ideal substitute for monolingual structure-based segmentation and only a rough approximation, as sentence structures in the German translations and structures in the DGS source text may differ to a some extent from each other.

For the purpose of this article we refer to two data types that are available to us: glosses and concepts.

4. Collocation

One of the advantages of a corpus is that collocational information can be extracted by statistical means from corpus data. Having a relatively large SL corpus available allows for new insights on sign usage and meaning and thus also for new kinds of information on signs in dictionary entries. In the context of our lexicographic work, collocational information is relevant in the following ways:

1. Collocational information is a good indicator of different word senses and can be utilised to support lexicographers in word sense discrimination (WSD).

2. Information on typical sign combinations and other usage patterns is information to be included in dictionary entries because it is useful especially for language learners. Lists of frequent neighbours of the target signs suggest candidates for collocations, phrases and (loan) compounds to be included in the dictionary entry of the *DW-DGS*.[6]

In this article we focus on methods of extracting collocational information from the *DGS Corpus* to support WSD. Considering that a) the investigation of collocations in sign language is only just becoming possible thanks to recent advances in SL corpus creation, b) DGS exhibits considerable phonological and lexical variation, and c) borders between individual signs and their variants might be less clear than their categorisation via glosses suggests, we adopt a rather pragmatic and broad definition of collocation for the time

[3] `http://meine-dgs.de`

[4] `http://ling.meine-dgs.de`

[5] For the purposes of this article we simplify the type structure slightly. For a discussion of the complete type structure, see Langer et al. (2018).

[6] `http://dw-dgs.meine-dgs.de`

left neighbour	base	right neighbour
MORE1	TIME1	FOR1
MORE1	TIME1	EQUAL1A
MORE1	TIME1	TO-SIGN1G
MORE1	TIME1	FOR1
MORE1	TIME1	TO-LOOK-AT1
MORE1	TIME1	$GEST-NM-NOD-HEAD1-$SAM
MORE3	TIME1	TO-CHANGE1B
MORE5	TIME1	TALK2A
MOTHER1	TIME1	$INDEX1
MUCH-OR-MANY1A	TIME1	BEAUTIFUL1A
MUCH-OR-MANY1A	TIME1	YOU1
MUCH-OR-MANY1A	TIME1	$GEST-OFF$GEST-OFF-$SAM
MUCH-OR-MANY1A	TIME1	FOR1
MUCH-OR-MANY1A	TIME1'phs:1	JOURNEY3
MUCH-OR-MANY1A	TIME1'phs:1	TO-DEDUCT2B

Figure 1: Excerpt of a list of 525 tokens of type `TIME1` with left and right neighbour glosses. Occurrences are sorted by the alphabetical order of the left neighbour.

left neighbour	base	right neighbour	pattern occ
I1	TIME1		28
	TIME1	I1	26
EQUAL1A	TIME1		14
BEAUTIFUL1A	TIME1		12
MORE1	TIME1		11
TO-NEED1	TIME1		11
I2	TIME1		10
	TIME1	FOR1	10
	TIME1	BARELY1	8
MUCH-OR-MANY1A	TIME1		7
	TIME1	TO-PRESSURE1	6
NONE3	TIME1		6
NONE1	TIME1		6
MY1	TIME1		6
GOOD1	TIME1		6
CERTAIN1	TIME1		5
	TIME1	WHATEVER3	5

Figure 2: Distinct neighbours, sorted by bi-gram frequency and a minimum frequency of five or more.

being.[7] In the context of lexicographic work with the dictionary user in mind the definition presented by Fuertes-Olivera et al. (2012, p. 299) can serve as a starting point: "The term collocation was chosen as an umbrella term for referring to word combinations that are typical for the kind of language in question, and which can be useful for reuse in text production or for assisting in text translation. They are composed of two or more orthographic words, do not constitute a full sentence, but offer potential users the possibility of obtaining relevant information [...]". In our case, we consider individual (simplex) signs represented by glosses, as our data contains no orthographic words.

Collocational information in our approach includes all kinds of multi-sign patterns, especially including *collocations* in the narrow sense and *selectional restrictions*. According to Atkins and Rundell (2008, p. 302), "[. . .] [b]oth terms refer to an observable tendency of certain words to occur frequently with certain other words. When we talk about 'selectional restrictions', we mean the general semantic category of items that typically appear as the subjects or objects of a verb, or or as the complements of an adjective. A collocation on the other hand, is a recurrent combination of words, where *one specific lexical item* (the 'node') has an observable tendency to occur with another (the 'collocate'), with a frequency greater than chance."[8]

Another definition of collocation that has "received general approval among lexicographers" (Orlandi, 2016, p. 26) is that of Bartsch (2004, p. 76) where collocations are "lexically and/or pragmatically constrained recurrent co-occurrences of at least two lexical items which are in a direct syntactical relation with each other". This definition includes what fully-fledged lexical profiles make explicit, that is, taking syntactic relations (e. g. dependencies) into account when determining and presenting collocations.

A relevant question for SL lexicography is how SL corpora can be used to automatically generate comparably informative collocation profiles for signs, even when many of the prerequesites for such an automated analysis do not yet exist for SL data. It would be very useful to find methods to identify frequent neighbours and to group them with regard to their syntactic or semantic relation to the base, be it argument structure or other kinds of functional or semantic categories. Collocation analysis for *DGS Corpus* data specifically is best done at the subtype level because subtypes correspond with conventional sign uses and pre-group occurrences according to these roughly defined meanings. In the following sections we discuss our approach to NLP supported detection, grouping, and presentation of collocations for lexicographic purposes.

5. Gloss-based Collocations

A very simple first approach for identifying frequent neighbours of a target sign, shown in Figure 1, is to run a query that returns all occurrences of the target sign and its left (or right) neighbours. Results are ordered alphabetically by the neighbour gloss name. The human eye will very quickly find groups of identical neighbours in this list.

The query can be refined to provide a better overview by showing each distinct left and right neighbour only once and count and display the number of occurrences for this combination (bi-gram) in the result. In Figure 2 we see a frequent neighbours list grouped by distinct neighbour glosses and ordered by the frequency count of the bi-gram. Groups with fewer than five members are filtered out.

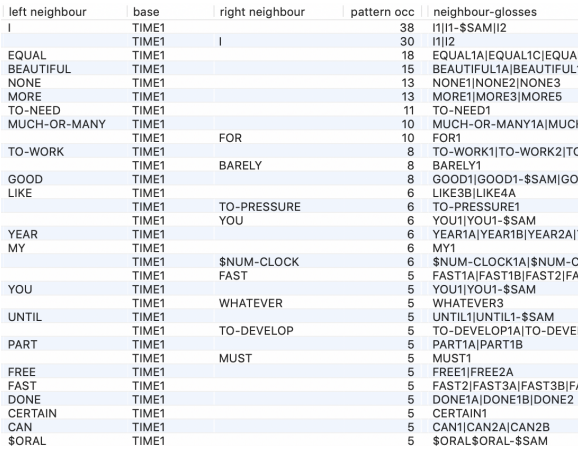
Up to this point the query shows collocations in the more narrow sense, i. e. frequent combinations of specific lexical items. However, as the corpus is still limited in size and most types have rather small numbers of tokens, relevant semantic patterns may not show up, especially since phonological variants are covered by separate types in the *DGS Corpus*.[9] For pattern detection the distinction of phonological variants is too fine-grained and variant types should be grouped together in the analysis. Furthermore, DGS not only exhibits a high amount of phonological variation, but also of lexical variation, often even within a single region or by a single individual.[10] For the purpose of WSD the

[7] For a good overview on defining collocation – especially in the field of lexicography – see Orlandi (2016).

[8] *Selectional restrictions* have also been called *selectional preference*, cf. Sinclair (1996).

[9] Phonological variants are marked by the same number but different letters after the gloss, e. g. `MUCH-OR-MANY1A` and `MUCH-OR-MANY1B`.

[10] In the *DGS Corpus* lexical variants share the same gloss name but receive different numbers, e. g. `NONE1` and `NONE3`. In some cases distinct meanings of a German word (polysemes and homonyms) may lead to the same gloss name being used for signs with distinct form and meaning. While this is a potential source of errors, it is in practice an acceptable trade-off, as such signs are expected to occur in distinctly different collocational contexts.



left neighbour	base	right neighbour	pattern occ	neighbour-glosses
I	TIME1		38	I1|I1-$SAM|I2
	TIME1	I	30	I1|I2
EQUAL	TIME1		18	EQUAL1A|EQUAL1C|EQUA
BEAUTIFUL	TIME1		15	BEAUTIFUL1A|BEAUTIFUL
NONE	TIME1		13	NONE1|NONE2|NONE3
MORE	TIME1		13	MORE1|MORE3|MORE5
TO-NEED	TIME1		11	TO-NEED1
MUCH-OR-MANY	TIME1		10	MUCH-OR-MANY1A|MUCI
	TIME1	FOR	10	FOR1
TO-WORK	TIME1		8	TO-WORK1|TO-WORK2|TO
	TIME1	BARELY	8	BARELY1
GOOD	TIME1		8	GOOD1|GOOD1-$SAM|GO
LIKE	TIME1		6	LIKE3B|LIKE4A
	TIME1	TO-PRESSURE	6	TO-PRESSURE1
	TIME1	YOU	6	YOU1|YOU1-$SAM
YEAR	TIME1		6	YEAR1A|YEAR1B|YEAR2A|
MY	TIME1		6	MY1
	TIME1	$NUM-CLOCK	6	$NUM-CLOCK1A\|$NUM-C
	TIME1	FAST	5	FAST1A|FAST1B|FAST2|FA
YOU	TIME1		5	YOU1|YOU1-$SAM
	TIME1	WHATEVER	5	WHATEVER3
UNTIL	TIME1		5	UNTIL1|UNTIL1-$SAM
	TIME1	TO-DEVELOP	5	TO-DEVELOP1A|TO-DEVE
PART	TIME1		5	PART1A|PART1B
	TIME1	MUST	5	MUST1
FREE	TIME1		5	FREE1|FREE2A
FAST	TIME1		5	FAST2|FAST3A|FAST3B|FA
DONE	TIME1		5	DONE1A|DONE1B|DONE2
CERTAIN	TIME1		5	CERTAIN1
CAN	TIME1		5	CAN1|CAN2A|CAN2B
$ORAL	TIME1		5	$ORAL$ORAL-$SAM

Figure 3: Frequent neighbours analysis similar to Figure 2, but with collapsed phonological and lexical variants. Variants included in a group are listed in its rightmost column.

focus of interest is less on specific lexical items (forms) but on the typical semantic context the target sign is used in. Grouping lexical as well as phonological variants together in neighbourhood pattern analysis can help to identify different senses of the specific target sign (base). Thus, for neighbourhood analysis not the individual types but all types with the same gloss name are collapsed into one group to leverage the information on phonological and lexical variation as coded in the full glosses. The result of this query can be seen in Figure 3.

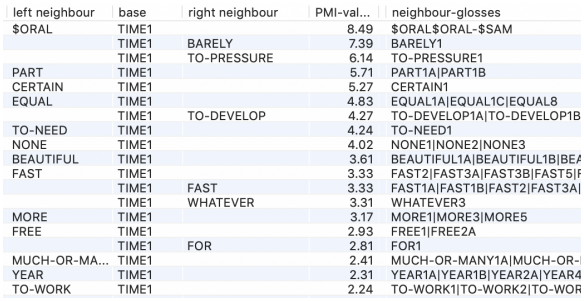
An advantage of this oversimplification is that more bi-gram combinations are feeding into the general semantic patterns so that more relevant patterns show up for the target sign. At this point the analysis is not covering individual collocations in the narrow sense anymore, but this level of granularity has proven fruitful for the corpus size and variant-richness of the *DGS Corpus*.

Regarding the sorting of results, there is further room for improvement. Sorting target-neighbour pairs by their raw co-occurrence frequency has an inherent bias towards signs that are generally more frequent, as they have a higher likelihood to co-occur with the target by chance without being particularly relevant. For example, in Figure 3 the gloss name `I` is ranked most highly because it covers two of the most frequent signs in the corpus, rather than due to any particular relevance for `TIME1`.

To address this bias, Church and Hanks (1990) introduced the use of **pointwise mutual information (PMI)** (Shannon, 1948) to lexicography.[11] Given the individual frequencies $f(x)$ and $f(y)$ of target and neighbour, their co-occurrence frequency $f(x, y)$ and the overall number of corpus tokens N, PMI is defined as:

$$\mathrm{I}(x, y) = \log_2 \frac{f(x, y)\, N}{f(x)\, f(y)} \tag{1}$$

[11] Lexicographers have since introduced a variety of other metrics, e. g. the *logDice* formula (Rychlý, 2008) used by *Sketch Engine*. Most of these require additional syntactic information and are therefore not suitable for our purposes (see Section 2.2).



left neighbour	base	right neighbour	PMI-val...	neighbour-glosses
$ORAL	TIME1		8.49	$ORAL$ORAL-$SAM
	TIME1	BARELY	7.39	BARELY1
	TIME1	TO-PRESSURE	6.14	TO-PRESSURE1
PART	TIME1		5.71	PART1A|PART1B
CERTAIN	TIME1		5.27	CERTAIN1
EQUAL	TIME1		4.83	EQUAL1A|EQUAL1C|EQUAL8
	TIME1	TO-DEVELOP	4.27	TO-DEVELOP1A|TO-DEVELOP1B
TO-NEED	TIME1		4.24	TO-NEED1
NONE	TIME1		4.02	NONE1|NONE2|NONE3
BEAUTIFUL	TIME1		3.61	BEAUTIFUL1A|BEAUTIFUL1B|BEA
FAST	TIME1		3.33	FAST2|FAST3A|FAST3B|FAST5|F
	TIME1	FAST	3.33	FAST1A|FAST1B|FAST2|FAST3A|
	TIME1	WHATEVER	3.31	WHATEVER3
MORE	TIME1		3.17	MORE1|MORE3|MORE5
FREE	TIME1		2.93	FREE1|FREE2A
	TIME1	FOR	2.81	FOR1
MUCH-OR-MA...	TIME1		2.41	MUCH-OR-MANY1A|MUCH-OR-
YEAR	TIME1		2.31	YEAR1A|YEAR1B|YEAR2A|YEAR4
TO-WORK	TIME1		2.24	TO-WORK1|TO-WORK2|TO-WOR

Figure 4: Collocation list view of frequent left and right neighbours of subtypes of the type `TIME1`, ordered by PMI value of the bi-gram combination; bi-grams with fewer than five occurrences are omitted.

Figure 4 shows the neighbourhood patterns query ordered by PMI measure. This collocation list view has proved itself useful for lexicographic analysis in the DGS-Korpus project for some years now and serves as a substitute for a not yet available full-grown collocational profile of the target sign under investigation.[12]

6. Supersense Collocations

The Word Sketch profiles enhance their collocation lists by providing a semantic clustering of collocates, e. g. by grouping near-synonyms together based on information from a thesaurus (cf. Atkins and Rundell, 2008, p. 111). In this section, we explore how semantic groupings can be realised for a sign language.

6.1. The Need for Semantic Categories

While using the collocation list view presented in Section 5 to support the analyses of sign usage, lexicographers in the DGS-Korpus project noticed wider semantic and syntactic patterns across listed neighbours. Sometimes several neighbours were identified as members of a category that could be described by an abstract criterion of semantic grouping or by a functional or presumed syntactic role. For example, several left neighbour collocates of `TIME1` are signs that have a gloss name indicating a quantifying relation with `TIME1`, e. g. `MUCH-OR-MANY`, `MORE` and `NONE`. Many left neighbour collocates of `TO-SAY1` are signs referring to persons filling the semantic role of agent as argument of `TO-SAY1`.

These patterns are often examples of semantic restriction/preference and at the same time can indicate dependency structures and syntactic functions, as these phenomena are often related (cf. Bartsch, 2004, pp. 70–71). They are very useful for WSD and also constitute valuable information on sign usage for language learners using the dictionary. Naturally, such gloss patterns must be verified against DGS data by inspecting the actual signed utterances.

However, some of these patterns may go unnoticed because each individual bi-gram contributing to the pattern may by itself be too infrequent to show up in the collocation list (see our use of frequency thresholds in Section 5). Only

[12] Langer et al. (2018) mention the approach, albeit in less detail, as part of their discussion of views for lexicographic work.

left supersense	left neighbour	base	PMI-value	pattern occ
Menge		**TIME1**	**0.96**	**61**
Menge	NONE	TIME1		13
Menge	MORE	TIME1		11
Menge	MUCH-OR-MANY	TIME1		10
Menge	FREE	TIME1		5
Menge	PART	TIME1		5
Menge	$NUM-CLOCK	TIME1		2
Menge	LITTLE-BIT	TIME1		2
Menge	PRESENT-OR-HERE	TIME1		2
Menge	$NUM-ONE-TO-TEN	TIME1		1
Menge	$NUM-TAPPING	TIME1		1
Menge	$NUM-TEEN	TIME1		1
Menge	$NUM-TENS	TIME1		1
Menge	$NUM-YEAR-AFTER-NOW	TIME1		1
Menge	$SPECIAL-NONE	TIME1		1
Menge	$SPECIAL-VERY	TIME1		1
Menge	ALL	TIME1		1
Menge	EVERYONE	TIME1		1
Menge	EVERYTHING	TIME1		1
Menge	OFTEN	TIME1		1
Attribut		**TIME1**	**0.73**	**27**
Attribut	EQUAL	TIME1		16
Attribut	FREE	TIME1		5
Attribut	STRICT	TIME1		2

Figure 5: Supersense collocation of TIME1. Supersenses are ranked by their PMI value. Below each supersense we show the gloss names that are part of its collocation.

left supersense	left neighbour	base	PMI-value	pattern occ
Lokation		**BACK1A**	**0.03**	**23**
Lokation	IT-WORKS-OUT	BACK1A		2
Lokation	TO-COME	BACK1A		2
Lokation	TO-DROP-OR-TO-GIVE-UP	BACK1A		2
Lokation	TO-FALL	BACK1A		2
Lokation	TO-GO	BACK1A		2
Lokation	AIR	BACK1A		1
Lokation	IT-HAPPENS	BACK1A		1
Lokation	TO-CLIMB	BACK1A		1
Lokation	TO-DRIVE	BACK1A		1
Lokation	TO-EAT-OR-FOOD	BACK1A		1
Lokation	TO-GET	BACK1A		1
Lokation	TO-GET-OUT	BACK1A		1
Lokation	TO-LAND	BACK1A		1
Lokation	TO-LET	BACK1A		1
Lokation	TO-SLIDE-OR-TO-PUSH	BACK1A		1
Lokation	TO-SWARM	BACK1A		1
Lokation	TO-WALK-AROUND	BACK1A		1
Lokation	TO-WASH-UP	BACK1A		1

Figure 6: Excerpt of the supersense collocation of BACK1, showing the supersense collocate Lokation (location) and the gloss names it contains. None of the individual names occurs more than twice in the corpus, but grouped into the supersense the semantic pattern becomes apparent.

after grouping all these infrequent signs together would it become apparent how frequent the pattern that they are part of may in fact be.

This presents us with a chicken and egg problem. To find the pattern we need to see the infrequent signs and to see the infrequent signs we need to have already grouped them according to our pattern. Syntactic information like parts of speech and dependency structures that might help structure our data further are not available to us. Instead, we take inspiration from Atkins and Rundell (2008), who mention *selectional restrictions* as manifestations of usage and therefore helpful to discriminate between different senses: "When we talk about 'selectional restrictions', we mean the general semantic category of items that typically appear as the subjects or objects of a verb, or as the complements of an adjective. [...] [O]nce you know the category, any word belonging to that category can fill the relevant slot" (Atkins and Rundell, 2008, pp. 302–303). While we cannot use selectional restrictions (this would require syntactic information about parts of speech and their arguments), we might still be able to group signs into semantic categories if we can access an appropriate semantic resource.

6.2. Wordnet Supersenses

A **wordnet** is a lexical resource of semantic relations between words of a specific language (Miller et al., 1990). Words are organised by their word senses and grouped with other words of the same sense into *synsets* (synonym sets). Synsets are linked by various relations, such as hyperonymy, meronymy or entailment. Each synset is also assigned a so-called *supersense* (also known as *lexicographer sense*). Supersenses are coarse semantic categories, such as person, location or emotion. They might therefore be used as semantic categories in our list of collocations.

No wordnet exists yet for DGS, so we instead leverage German-language components of the *DGS Corpus* to extract supersenses from the German wordnet *GermaNet* (Hamp and Feldweg, 1997). As we are looking to retrieve semantic generalisations, rather than nuances, we believe this to be an acceptable compromise.

To connect DGS signs to *GermaNet* supersenses, we use the concept entries associated with subtypes in the *DGS Corpus* (see Section 3.2). As concept entries are (approximate) indications of the conventional meanings of a sign and are written as single German expressions, we treat them as rough German equivalents. Each sign can have several concepts, giving us a one-to-many mapping to German words. For each sign concept we look up matching terms in *GermaNet* across all parts of speech and retrieve the synsets which they are part of. The supersenses of these synsets are then treated as the supersenses of the concept. The supersenses of a sign are the set of supersenses of all concepts of that sign. Similarly, the supersenses of a gloss name group (i. e. a set of signs grouped by the name-component of their gloss, see Section 5) consist of all supersenses of its signs.

Having now bootstrapped supersense categories for our DGS sign inventory, we return to the task of creating a collocation list view. First we follow the steps of the gloss-based collocations pipeline from Section 5. The neighbouring tokens of the base are grouped by their signs to establish collocates. These sign collocates are collapsed further into gloss name collocates. Instead of then listing the gloss name collocates directly, we look up their supersenses and use them to create supersense collocates. Each supersense collocate contains every associated gloss name collocate. This means a gloss name collocate can occur in multiple supersense collocates if it has multiple supersenses. The supersense collocates are then ranked by their PMI and the list is pruned at the usual frequency threshold. In the collocation list view, each supersense collocate is followed by a list of the gloss name collocates that it is comprised of. An example of this can be seen in Figure 5.

Note that the frequency threshold only applies to the supersense collocate, not to individual gloss name collocates it contains. This allows the long tail of low-frequency collocations to still impact the ranking of the semantic categories that they are a part of. For example, Figure 6 shows an excerpt of the supersense collocation of BACK1. One sense of this sign can be described as *"moving back to a place where one came from or has been before"*. Often this sign follows other signs of movement across space, such as TO-GO-THERE, TO-COME, TO-DRIVE, TO-GET-OUT and others. However, none of these neighbour signs co-

occur with BACK1 more than once or twice in the corpus. Due to these low individual frequencies, they are omitted in the gloss name collocation view. As we can see in Figure 6, the same is not true for the supersense collocation view. Here the neighbour signs are grouped together under the supersense Lokation (location), clearly showing the collocation pattern of the sign sense. Cases like this show how the supersense collocation view can be a useful addition to the lexicographer's toolkit, especially when used in concert with the gloss name collocation view.

6.3. Fallback: Glosses as Concepts

Our approach for connecting signs with *GermaNet* supersenses relies on the availability of concept entries as a bridge between languages. However, for many other sign language datasets, such an explicit cross-lingual semantic layer is not available. A possible fallback solution can be the use of gloss names as impromptu concepts. While there are obvious drawbacks to this (no multiple concepts per sign, as well as the established dangers of treating glosses as translations) it may still be an acceptable compromise to provide lexicographers with another tool in their toolbox.

On the technical side, certain complications arise as well. As glosses are commonly written in all caps, capitalisation of the host language is lost, which may create ambiguities, depending on the language (e. g. in German *laut* means 'loud', but *Laut* means 'sound'). Depending on the exact annotation guidelines used for naming glosses, a variety of multi-word gloss ambiguities may also have to be resolved. In the *DGS Corpus*, for example, hyphenation can fulfil a number of different functions. It can indicate actual multi-word expressions (ACH-SO1, *ach so*, 'I see') or fine-grained meanings that require more than a single German term to describe (ANMACHEN-BILDSCHIRM1, *anmachen (Bildschirm)*, 'turn on (monitor)'). It can be used to provide disambiguating contexts (FREI-KOSTENLOS1 means 'free' (*frei*) in the sense of 'at no charge' (*kostenlos*), but not 'unclaimed' or 'not imprisoned') or to append foreign language markers (HOLLYWOOD-ASL1 is a sign in American Sign Language). And, of course, sometimes a hyphen is simply part of a word (S-BAHN1, *S-Bahn*, 'commuter train').

6.4. Caveats and Finetuning

Using *GermaNet* supersenses to group signs is a first step towards bootstrapping semantic categories for DGS. It is not, however, without its caveats. The first one is that supersenses are extremely broad categories. Not counting duplicates across different parts of speech, *GermaNet* provides a total of 38 different supersenses. Supersense collocations can therefore only ever be a first filter, followed by a more thorough analysis.

Another problem stems from the fact that we access *GermaNet*, a primarily sense-based resource, via lemma-based lookups. As we are unable to determine ahead of time which word senses apply to the tokens in a given collocate, we are bound to overgenerate, selecting more senses than necessary and thus extracting incorrect supersenses. This issue is exacerbated by the fact that we are performing this lookup cross-lingually, thus capturing word senses of a German translation that do not apply to the sign at all. We expect that the impact of these issues is lessened in our specific case, as many of the incorrect senses will share a supersense with correct senses. Also, as the resulting output is intended for lexicographic work, any suggested patterns will be further scrutinised by the lexicographer.

The third issue is one of lexical coverage. While *GermaNet* covers a very large vocabulary, it focuses on content words, especially nouns, verbs and adjectives. Function words are omitted. It also does not cover names and has only a limited selection of location names.

To address the last two issues at least in part, we introduce a number of additional steps when determining supersenses for signs. As was mentioned in Section 3.2, concept entries can be given a disambiguating *context* when the German concept term by itself is too ambiguous (e. g. *"Bayern"* can refer to either the German federal state or a football club). While the context field generally contains freeform text, certain contexts occur repeatedly (e. g. *"Ortsname"* (place name) for city names). Such contexts can be used to assign supersenses (or other semantic categories) directly without having to consult *GermaNet*. Similarly, certain glosses have semantic prefixes that can be used directly, such as the $NAME- prefix for person names or $NUM- prefix for numbers and related terms. Finally, we introduce a pseudo-supersense called stopwords to which we assign signs whose German context word is found in a list of common stopwords.

Using these finetuning steps in concert with the pipeline described in Section 6.2, we are able to assign supersenses to 94% of *DGS Corpus* subtypes. Of these, 82% are assigned three supersenses or less and 46% are assigned a single one.

7. Outlook

In this article we presented approaches for creating collocation views for sign language research by using glosses and wordnet supersenses. In the future we hope to improve upon these in several ways. For example, up until now we only consider immediate neighbours of a target sign. However, collocates can also be separated from the target by other signs, so future collocation analyses should consider larger windows or skip-grams (cf. Järvelin et al., 2007). We also envisage dynamic merging of the right and left neighbour lists in cases where collocates seem to follow a free word order.

We hope to extend our cross-lingual bootstrapping of semantic information to finer semantic information than supersenses, such as near-synonyms or the hyperonymy hierarchy of *GermaNet*. This introduces new challenges, such as how to group terms within the hierarchy, and increases the relevance of known issues, like the overgeneralisation we face when selecting word senses.

Another exciting possibility is the potential of creating a feedback loop between the cross-lingual bootstrapping of wordnet information and the word sense discrimination performed by the lexicographers. While we showed how the bootstrapping can help lexicographers, we hope that in return the lexicographers' descriptions can improve the bootstrapping process by providing sign sense inventories and select token sense tags.

8. Acknowledgements

This publication has been produced in the context of the joint research funding of the German Federal Government and Federal States in the Academies' Programme, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies' Programme is coordinated by the Union of the Academies of Sciences and Humanities.

9. Bibliographical References

Atkins, B. T. S. and Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford University Press, New York, USA.

Auer, P. (2010). Zum Segmentierungsproblem in der Gesprochenen Sprache. Pre-Publication in *InLiSt – Interaction and Linguistic Structures*, Number 49.

Bartsch, S. (2004). *Structural and Functional Properties of Collocations in English: A Corpus Study of Lexical and Pragmatic Constraints on Lexical Co-Occurrence*. Doctoral thesis, Tübingen : Naar, Darmstadt, Germany.

Church, K. W. and Hanks, P. (1990). Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16(1):22–29, March.

Fuertes-Olivera, P. A., Bergenholtz, H., Nielsen, S., and Niño Amo, M. (2012). Classification in Lexicography: The Concept of Collocation in the Accounting Dictionaries. *Lexicographica*, 28(1):293–308.

Hamaker, J., Zeng, Y., and Picone, J. (1998). Rules and Guidelines for Transcription and Segmentation of the SWITCHBOARD Large Vocabulary Conversational Speech Recognition Corpus. Technical report version 7.1, Institute for Signal and Information Processing, Mississippi State University, Starkville, Mississippi, USA.

Hamp, B. and Feldweg, H. (1997). GermaNet - a Lexical-Semantic Net for German. In *Proceedings of the ACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15, Madrid, Spain. ACL.

Hanke, T. and Storz, J. (2008). iLex – A Database Tool for Integrating Sign Language Corpus Linguistics and Sign Language Lexicography. In *Proceedings of SignLang@LREC*, pages 64–67, Marrakech, Morocco. ELRA.

Hanke, T. (2004). Hamnosys – Representing Sign Language Data in Language Resources and Language Processing Contexts. In *Proceedings of SignLang@LREC*, pages 1–6, Lisbon, Portugal. ELRA.

Järvelin, A., Järvelin, A., and Järvelin, K. (2007). s-grams: Defining generalized n-grams for information retrieval. *Information Processing & Management*, 43(4):1005–1019.

Johnston, T. (2010). From Archive to Corpus: Transcription and Annotation in the Creation of Signed Language Corpora. *International Journal of Corpus Linguistics*, 15(1):106–131.

Konrad, R., Hanke, T., Langer, G., König, S., König, L., Nishio, R., and Regen, A. (2018). Public DGS Corpus: Annotation Conventions. Project Note AP03-2018-01, DGS-Korpus project, IDGS, Hamburg University, Hamburg, Germany.

Landau, S. I. (2001). *Dictionaries: The Art and Craft of Lexicography*. Cambridge University Press, 2nd edition.

Langer, G., Troelsgård, T., Kristoffersen, J., Konrad, R., Hanke, T., and König, S. (2016). Designing a Lexical Database for a Combined Use of Corpus Annotation and Dictionary Editing. In *Proceedings of SignLang@LREC*, pages 143–152, Portorož, Slovenia. ELRA.

Langer, G., Müller, A., and Wähl, S. (2018). Queries and Views in iLex to Support Corpus-Based Lexicographic Work on German Sign Language (DGS). In *Proceedings of SignLang@LREC*, pages 107–114, Miyazaki, Japan. ELRA.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to WordNet: An Online Lexical Database. *International Journal of Lexicography*, 3:235–244.

Nishio, R., Hong, S.-E., König, S., Konrad, R., Langer, G., Hanke, T., and Rathmann, C. (2010). Elicitation Methods in the DGS (German Sign Language) Corpus Project. In *Proceedings of SignLang@LREC*, pages 178–185, Valletta, Malta. ELRA.

Orlandi, A. (2016). Monolingual collocation lexicography: State of art and new perspectives. In Adriana Orlandi et al., editors, *Defining collocation for lexicographic purposes – From linguistic theory to lexicographic practice*, number 219 in Linguistic Insights, pages 19–70. Peter Lang, Bern, Switzerland.

Rychlý, P. (2008). A Lexicographer-Friendly Association Score. In *Proceedings of RASLAN*, pages 6–9, Karlova Studánka, Czech Republic.

Schmidt, T. (2014). The Research and Teaching Corpus of Spoken German — folk. In *Proceedings of LREC*, pages 383–387, Reykjavik, Iceland. ELRA).

Schwager, W. and Zeshan, U. (2008). Word Classes in Sign Languages: Criteria and Classifications. *Studies in Language*, 32(3):509–545.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27(3):379–423.

Sinclair, J. M. (1996). The search for units of meaning. *Textus*, 9(1):75–106.

Slobin, D. I. (2008). Breaking the Molds: Signed Languages and the Nature of Human Language. *Sign Language Studies*, 8(2):114–130.

Westpfahl, S. and Gorisch, J. (2018). A Syntax-Based Scheme for the Annotation and Segmentation of German Spoken Language Interactions. In *Proceedings of LAW-MWE-CxG 2018@COLING*, pages 109–120, Santa Fe, New Mexico, USA. ACL.

Zwitserlood, I. E. P., Kristoffersen, J. H., Troelsgård, T., and Jackson, H. (2013). Issues in sign language lexicography. In *Bloomsbury Companion To Lexicography*, Bloomsbury Companions, pages 259–283. Bloomsbury Continuum, London, United Kingdom.

Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 135–138
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

Machine Learning for Enhancing Dementia Screening in Ageing Deaf Signers of British Sign Language

Xing Liang, Bencie Woll, Epaminondas Kapetanios, Anastasia Angelopoulou, Reda Al batat
IoT and Security Research Group, University of Greenwich, UK
Cognitive Computing Research Lab, University of Westminster, UK
Deafness Cognition and Language Research Centre, University College London, UK
x.liang@greenwich.ac.uk, b.woll@ucl.ac.uk, (kapetae, agelopa)@westminster.ac.uk, w1601767@my.westminster.ac.uk

Abstract

Real-time hand movement trajectory tracking based on machine learning approaches may assist the early identification of dementia in ageing deaf individuals who are users of British Sign Language (BSL), since there are few clinicians with appropriate communication skills, and a shortage of sign language interpreters. In this paper, we introduce an automatic dementia screening system for ageing Deaf signers of BSL, using a Convolutional Neural Network (CNN) to analyse the sign space envelope and facial expression of BSL signers recorded in normal 2D videos from the BSL corpus. Our approach involves the introduction of a sub-network (the multi-modal feature extractor) which includes an accurate real-time hand trajectory tracking model and a real-time landmark facial motion analysis model. The experiments show the effectiveness of our deep learning based approach in terms of sign space tracking, facial motion tracking and early stage dementia performance assessment tasks.

Keywords: Real-Time Hand Tracking, Facial Analysis, British Sign Language, Dementia, Convolutional Neural Network

1. Introduction

British Sign Language (BSL), is a natural human language, which, like other sign languages, uses movements of the hands, body and face for linguistic expression. Identifying dementia in BSL users, however, is still an open research field, since there is very little information available about the incidence or features of dementia among BSL users. This is also exacerbated by the fact that there are few clinicians with appropriate communication skills and experience working with the BSL-using population. Diagnosis of dementia is subject to the quality of cognitive tests and BSL interpreters alike. Hence, the Deaf community currently receives unequal access to diagnosis and care for acquired neurological impairments, with consequent poorer outcomes and increased care costs (Atkinson et al., 2002). In this context, we propose a methodological approach to initial screening that comprises several stages. The first stage of research focuses on analysing the motion patterns of the sign space envelope in terms of sign trajectory and sign speed by deploying a real-time hand movement trajectory tracking model (Liang et al., 2019) based on OpenPose[1] library. The second stage involves the extraction of the facial expressions of deaf signers by deploying a real-time facial analysis model based on dlib library[2] to identify active and non-active facial expressions. Based on the differences in patterns obtained from facial and trajectory motion data, the further stage of research implements both VGG16 (Simonyan and Zisserman, 2015) and ResNet-50 (He et al., 2016) networks using transfer learning from image recognition tasks to incrementally identify and improve recognition rates for Mild Cognitive Impairment (MCI) (i.e. pre-dementia). Performance evaluation of the research work is based on data sets available from the Deafness Cognition and Language Research Centre (DCAL) at UCL, which has a range of video recordings of over 500 signers who have volunteered to participate in research. Figure 1 shows the pipeline and high-level overview of the network design.

The paper is structured as follows: Section 2 gives an overview of the related work. Section 3 outlines the methodological approach followed by Section 4 with the discussion of experimental design and results. A conclusion provides a summary of the key contributions and results of this paper.

2. Related Work

Recent advances in computer vision and greater availability in medical imaging with improved quality have increased the opportunities to develop machine learning approaches for automated detection and quantification of diseases, such as Alzheimer's and dementia (Pellegrini et al., 2018). Many of these techniques have been applied to the classification of MR imaging, CT scan imaging, FDG-PET scan imaging or the combined imaging of above, by comparing patients with early stage disease to healthy controls, to distinguish different types or stages of disease and accelerated features of ageing (Spasova et al., 2019; Lu et al., 2018; Huang et al., 2019). In terms of dementia diagnosis (Astell et al., 2019), there have been increasing applications of various machine learning approaches, most commonly with imaging data for diagnosis and disease progression (Negin et al., 2018; Iizuka et al., 2019) and less frequently in non-imaging studies focused on demographic data, cognitive measures (Bhagyashree et al., 2018), and unobtrusive monitoring of gait patterns over time (Dodge et al., 2012). These and other real-time measures of function may offer novel ways of detecting transition phases leading to dementia, which could be another potential research extension to our toolkit, since the real-time hand trajectory tracking submodel has the potential to track a patient's daily walking

[1] https://github.com/CMU-Perceptual-Computing-Lab/openpose
[2] http://dlib.net/

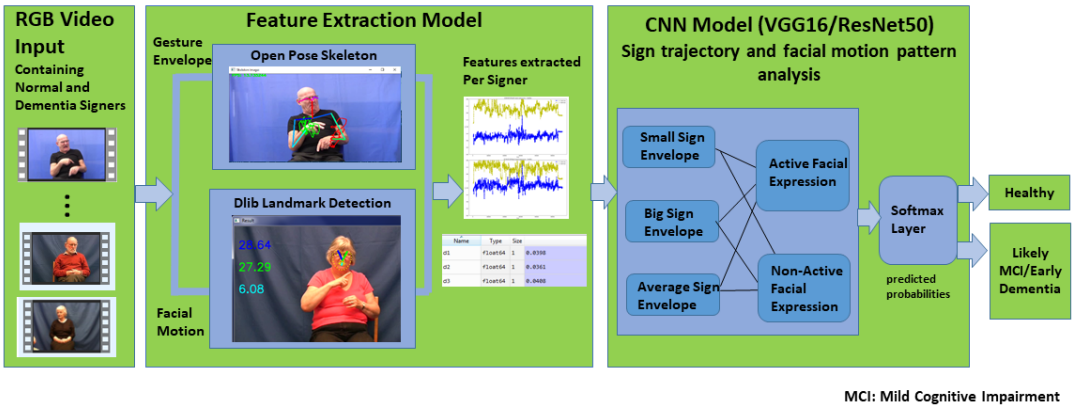


Figure 1: The Proposed Pipeline for Dementia Screening

pattern and pose recognition as well.

3. Methodology

In this paper, we present a multi-modal feature extraction sub-network inspired by practical clinical needs, together with the experimental findings associated with the sub-network. The input to the system is short term clipped videos. Different extracted motion features are fed into the CNN network to classify a BSL signer as healthy or atypical. Performance evaluation of the research work is based on data sets available from the BSL Corpus[3] at DCAL UCL, a collection of 2D video clips of 250 Deaf signers of BSL from 8 regions of the UK; and two additional data sets: a set of data collected for a previous funded project [4], and a set of signer data collected for the present study.

3.1. Dataset

From the video recordings, we selected 40 case studies of signers (20M, 20F) aged between 60 and 90 years; 21 are signers considered to be healthy cases based on their scores on the British Sign Language Cognitive Screen (BSL-CS); 9 are signers identified as having Mild Cognitive Impairment (MCI) on the basis of the BSL-CS; and 10 are signers diagnosed with MCI through clinical assessment. We consider those 19 cases as MCI (i.e. early dementia) cases, whether identified through the BSL-CS or clinically. As the video clip for each case is about 20 minutes in length, we segmented each into 4-5 short video clips - 4 minutes in length - and fed the segmented short video clip to the multi-modal feature extraction sub-network. In this way, we were able to increase the size of the dataset from 40 to 162 clips. Of the 162, 79 have MCI, and 83 are cognitively healthy.

3.2. Real-time Hand Trajectory Tracking Model

OpenPose, developed by Carnegie Mellon University, is one of the state-of-the-art methods for human pose estimation, processing images through a 2-branch multi-stage CNN (Cao et al., 2017). The real-time hand movement

[3]BSL Corpus Project, https://bslcorpusproject.org/.

[4]Overcoming obstacles to the early identification of dementia in the signing Deaf community

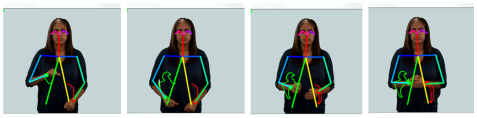


Figure 2: Real-Time Hand Trajectory Tracking for the Sign FARM

trajectory tracking model is developed based on the OpenPose Mobilenet Thin model (OpenPoseTensorFlow, 2019). A detailed evaluation of tracking performance is discussed in (Liang et al., 2019). The inputs to the system are brief clipped videos. We assume that the subjects are in front of the camera with only the head, upper body and arms visible; therefore, only 14 upper body parts in the image are outputted from the tracking model. These are: eyes, nose, ears, neck, shoulders, elbows, wrists, and hips. The hand movement trajectory is obtained via wrist joint motion trajectories. The curve of the hand movement trajectory is connected by the location of the wrist joint keypoints to track left- and right-hand limb movements across sequential video frames in a rapid and unique way. Figure 2, demonstrates the tracking process for the sign FARM. As shown in Figure 3, left- and right-hand trajectories obtained from the tracking model are also plotted by wrist location X and Y coordinates over time in a 2D plot. Figure 3 shows how hand motion changes over time, which gives a clear indication of hand movement speed (X-axis speed based on 2D coordinate changes, and Y-axis speed based on 2D coordinate changes). A spiky trajectory indicates more changes within a shorter period, thus faster hand movement.

3.3. Real-time Facial Analysis Model

The facial analysis model was implemented based on a facial landmark detector inside the Dlib library, in order to analyse a signer's facial expressions (Kazemi and Sullivan, 2014). The face detector uses the classic Histogram of Oriented Gradients (HOG) feature combined with a linear classifier, an image pyramid, and a sliding window detection scheme. The pre-trained facial landmark detector is used

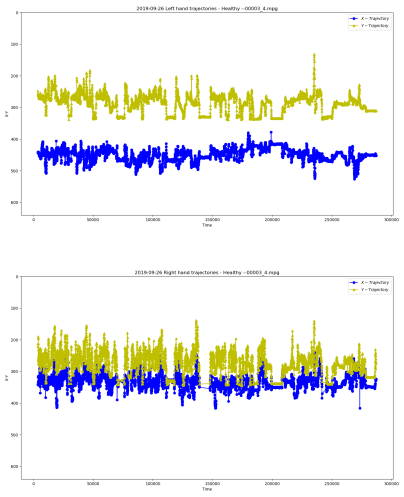


Figure 3: 2D Left- and Right- Hand Trajectory of a Signer

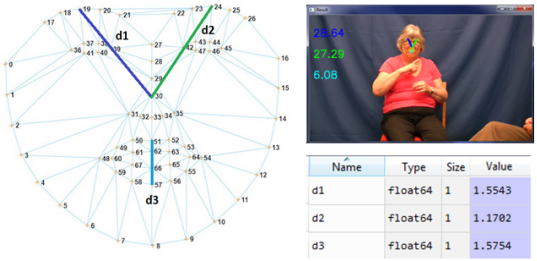


Figure 4: Facial Motion Tracking of a Signer

to estimate the location of 68 (x,y) coordinates that map to facial features (Figure 4). The facial analysis model extracts subtle facial muscle movement by calculating the average Euclidean distance differences between the nose and right brow as d1, nose and left brow as d2,and upper and lower lips as d3 for a given signer over a sequence of video frames (Figure 4). The vector [d1, d2, d3] is an indicator of a signer's facial expression and is used to classify a signer as having an active or non-active facial expression.

$$d1, d2, d3 = \frac{\sum_{t=1}^{T} |d^{t+1} - d^{t}|}{T} \quad (1)$$

where T = Total number of frames that facial landmarks are detected.

4. Experiments and Analysis

4.1. Experiments

In our approach, we have used VGG16 and ResNet-50 as the base models, with transfer learning to transfer the parameters pre-trained for the 1000 object detection task on the ImageNet dataset to recognise hand movement trajectory images for early MCI screening. We run the experiments on a Windows desktop with two Nvidia GeForce GTX 1080Ti adapter cards and 3.3 GHz Intel Core i9-7900X CPU with 16 GB RAM. In the training process, videos of 40 participants have been segmented into short clips with 162 segmented cases, split into 80% for the training set and 20% for the test set. To validate the model performance, we also kept 6 cases separate (1 MCI and 5 healthy signers), segmented into 24 cases for performance validation. Due to the very small dataset, we train ResNet-50 as a classifier alone and fine tune the VGG 16 network by freezing the Convolutional (Covn) layers and two Fully Connected (FC) layers, and only retrain the last two layers. Subsequently, a softmax layer for binary classification is applied to discriminate the two labels: Healthy and MCI, producing two numerical values of which the sum becomes 1.0. During training, dropout was deployed in fully connected layers and EarlyStopping was used in both networks to avoid overfitting.

4.2. Results Discussion

During test and validation, accuracies and receiver operating characteristic (ROC) curves of the classification were calculated, and the network with the highest accuracy and area under ROC (AUC), that is VGG 16, was chosen as the final classifier. Table 1 summarises the results over 46 participants from both networks. The best performance metrics are achieved by VGG16 with test set accuracy of 87.8788%, which matches validation set accuracy of 87.5%. In Figure 5, feature extraction results show that in a greater number of cases a signer with MCI produces a sign trajectory that resembles a straight line rather than the spiky trajectory characteristic of a healthy signer. In other words, signers with MCI produced more static poses/pauses during signing, with a reduced sign space envelope as indicated by smaller amplitude differences between the top and bottom peaks of the X, Y trajectory lines. At the same time, the Euclidean distance d3 of healthy signers is larger than that of MCI signers, indicating active facial movements by healthy signers. This proves the clinical observation concept of differences between signers with MCI and healthy signers in the envelope of sign space and face movements, with the former using smaller sign space and limited facial expression.

5. Conclusions

We have outlined a methodological approach and developed a toolkit for an automatic dementia screening system for signers of BSL. As part of our methodology, we report the experimental findings for the multi-modal feature extractor sub-network in terms of sign trajectory and facial motion together with performance comparisons between different CNN models in ResNet-50 and VGG16. The experiments show the effectiveness of our deep learning based approach for early stage dementia screening. The results are validated against cognitive assessment scores with a test set performance of 87.88%, and a validation set performance of 87.5% over sub-cases.

Astell, A., Bouranis, N., Hoey, J., Lindauer, A., Mihailidis, A., Nugent, C., and Robillardi, J. (2019). Technology and dementia: The future is now. *In: Dementia and Geriatric Cognitive Disorders*, 47(3):131–139.

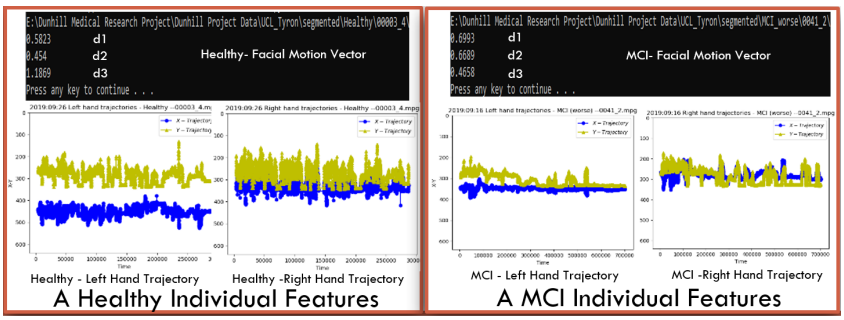


Figure 5: Experiment Finding

Table 1: Performance Evaluation over VGG16 and RestNet-50 for early MCI screening

Method	40 Participants 21 Healthy, 19 Early MCI			6 Participants 5 Healthy, 1 Early MCI	
	Train Result (129 segmented cases)	Test Result (33 segmented cases)		Validation Result (24 segmented cases)	
	ACC	ACC	ROC	ACC	ROC
VGG 16	87.5969%	**87.8788%**	**0.93**	**87.5%**	**0.96**
ResNet-50	69.7674%	69.6970%	0.72	66.6667%	0.73

Atkinson, J., Marshall, J., Thacker, A., and Woll, B. (2002). When sign language breaks down: Deaf people's access to language therapy in the uk. *In: Deaf Worlds*, 18:9–21.

Bhagyashree, S. I., Nagaraj, K., Prince, M., Fall, C., and Krishna, M. (2018). Diagnosis of dementia by machine learning methods in epidemiological studies: a pilot exploratory study from south india. *In: Social Psychiatry and Psychiatric Epidemiology*, 53(1):77–86.

Cao, Z., Simon, T., Wei, S., and Sheikh, Y. (2017). Real-time multi-person 2d pose estimation using part affinity fields. *In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dodge, H., Mattek, N., Austin, D., Hayes, T., and Kaye, J. (2012). In-home walking speeds and variability trajectories associated with mild cognitive impairment. *In: Neurology*, 78(24):1946–1952.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. *In: Proceedings of Computer Vision and Pattern Recognition (CVPR)*.

Huang, Y., Xu, J., Zhou, Y., Tong, T., Zhuang, X., and ADNI. (2019). Diagnosis of alzheimer's disease via multi-modality 3d convolutional neural network. *In: Front Neuroscience*, 13(509).

Iizuka, T., Fukasawa, M., and Kameyama, M. (2019). Deep-learning-based imaging-classification identified cingulate island sign in dementia with lewy bodies. *In: Scientific Reports*, 9(8944).

Kazemi, V. and Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. *In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liang, X., Kapetanios, E., Woll, B., and Angelopoulou, A. (2019). Real Time Hand Movement Trajectory Tracking for Enhancing Dementia Screening in Ageing Deaf Signers of British Sign Language. *Cross Domain Conference for Machine Learning and Knowledge Extraction (CD-MAKE 2019). Lecture Notes in Computer Science*, 11713:377–394.

Lu, D., Popuri, K., Ding, G. W., Balachandar, R., Beg, M., and ADNI. (2018). Multimodal and multiscale deep neural networks for the early diagnosis of alzheimer's disease using structural mr and fdg-pet images. *In: Scientific Reports*, 8(1):5697.

Negin, F., Rodriguez, P., Koperski, M., Kerboua, A., Gonzàlez, J., Bourgeois, J., Chapoulie, E., Robert, P., and Bremond, F. (2018). Praxis: Towards automatic cognitive assessment using gesture. *In: Expert Systems with Applications*, 106:21–35.

OpenPoseTensorFlow. (2019). https://github.com/ildoonet/tf-pose-estimation.

Pellegrini, E., Ballerini, L., Hernandez, M., Chappell, F., González-Castro, V., Anblagan, D., Danso, S., Maniega, S., Job, D., Pernet, C., Mair, G., MacGillivray, T., Trucco, E., and Wardlaw, J. (2018). Machine learning of neuroimaging to diagnose cognitive impairment and dementia: a systematic review and comparative analysis. *In: Alzheimer's Dementia: Diagnosis, Assessment Disease Monitoring*, 10:519–535.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional net-works for large-scale image recognition. *In: Proceedings of International Conference on Learning Representations*.

Spasova, S., Passamonti, L., Duggento, A., Liò, P., Toschi, N., and ADNI. (2019). A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to alzheimer's disease. *In: NeuroImage*, 189:276–287.

Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 139–144
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

Machine Translation from Spoken Language to Sign Language using Pre-trained Language Model as Encoder

Taro Miyazaki, Yusuke Morita, Masanori Sano
NHK Science and Technology Research Laboratories
1-10-11 Kinuta, Setagaya-ku, Tokyo, Japan
{miyazaki.t-jw, morita.y-gm, sano.m-fo}@nhk.or.jp

Abstract

Sign language is the first language for those who were born deaf or lost their hearing in early childhood, so such individuals require services provided with sign language. To achieve flexible open-domain services with sign language, machine translations into sign language are needed. Machine translations generally require large-scale training corpora, but there are only small corpora for sign language. To overcome this data-shortage scenario, we developed a method that involves using a pre-trained language model of spoken language as the initial model of the encoder of the machine translation model. We evaluated our method by comparing it to baseline methods, including phrase-based machine translation, using only 130,000 phrase pairs of training data. Our method outperformed the baseline method, and we found that one of the reasons of translation error is from *Pointing*, which is a special feature used in sign language. We also conducted trials to improve the translation quality for *Pointing*. The results are somewhat disappointing, so we believe that there is still room for improving translation quality, especially for *Pointing*.

Keywords: Japanese Sign Language, Machine Translation, BERT, Pointing

1. Introduction

Sign language is the first language for those who were born deaf or lost their hearing in early childhood. Such individuals understand sign language better than transcribed spoken language because sign languages differ from spoken languages in not only the modals to express meanings but also in grammar and vocabulary. Therefore, they require services provided with sign language, but there are only a few services provided. For example, less than 0.5% of airtime of TV programs have sign language services in Japan (Ministry of Internal Affairs and Communications, 2019).

There are many efforts to develop systems to provide more services in sign language through computer graphics (CG)-based animation (Kipp et al., 2011; Romeo et al., 2014; Uchida et al., 2018; Azuma et al., 2018). These systems are designed for practical domain-specific services. Therefore, they apply rule-based translation methods. Typical rule-based translation methods can translate with high quality for the target domain, but the coverage for the input tends to be narrow.

To provide sign language services that can be used for flexible, open-domain target contents, non-rule-based machine translation is necessary. Machine translations generally require large-scale training corpora (Koehn and Knowles, 2017; Lample et al., 2018). However, there are only small corpora for sign languages; one reason is that sign languages do not have writing systems.

To overcome this data-shortage scenario, we use a pre-trained language model of spoken language for the machine translation model. Our method is based on Transformer (Vaswani et al., 2017), and we use BERT (Devlin et al., 2019) as the initial model of the encoder. The encoder embeds the input transcribed spoken language, then the embedded vectors are fed to the decoder, which is also based on Transformer, then the input sentences are translated into sign language glosses. Evaluation results indicate that our method outperformed baseline methods, including phrase-based statistical machine translation (PBSMT)-based method, using only 130,000 sentence pairs of training data.

We also show that one of the reasons of translation error is from *Pointing*, which is typically used as pronoun (Cormier et al., 2013). Thus, we also conducted trials of accurately translating *Pointing*.

Our contributions are as follows: (1) We apply Transformer-based neural machine translation (NMT) from spoken language to sign language by using a pre-trained language model as the initial model of the encoder of the translation model,, (2) This method outperformed baseline methods, including PBSMT with training data of 130,000 sentence pairs, which is a small amount of training data for NMT, (3) We share our experiences of a trial to improve translation quality for *Pointing*, the results of which are somewhat disappointing, but include important suggestions.

2. Related Work

2.1. Sign Language Translation

Statistical machine translation (SMT) methods are widely used, so many studies on sign language translation are based on such methods. Stein et al. (2010) applied many SMT techniques and obtained high translation quality with a small corpus. San-Segundo et al. (2012) reported on combining three translation methods — example-based, rule-based, and SMT — to translate from spoken Spanish to sign language.

There are several methods that adopt special features of sign language such as mouthing, facial expression, and expression speed. Massó and Badia (2010) used these special features for training data and obtain good results. Morrissey (2011) used HamNoSys (Hanke, 2004) as a sign language translation method, which can be expanded by taking into account not only the word meanings but also facial and other expressions.

NMT is currently the mainstream in machine translation research. However, not many studies apply NMT for sign language translation because NMT methods require much more training data than SMT methods. Mocialov et al. (2018) showed that transfer learning is effective in improving the perplexity of long short-term memory (LSTM)-based language models for sign language. Stoll et al. (2018) used an encoder-decoder-based NMT method in their end-to-end spoken language to sign language video translation system. Cihan Camgoz et al. (2018) proposed an attention-based encoder-decoder translation method from sign video to spoken language by comparing various methods of embedding, tokenizing etc. Most NMT-based sign language translation methods use domain-specific data. Therefore, the translation quality for the domain is high, but the coverage for the input is narrow because the vocabulary size for sign language is small (around 1,000). Our model has a vocabulary size of 6,000, which differs from those used in prior studies.

2.2. Low-resource Languages Translations

There have been many studies on translating from/into low-resource languages, which are also very informative for improving machine translation of sign language because sign languages are also low-resource languages.

Dabre et al. (2019) proposed a technique of transfer learning based on multistage fine-tuning between small multi-parallel corpora to train a one-to-many NMT model. Skorokhodov et al. (2018) proposed an approach of initializing a translation model with language models. These two studies are based on transfer learning, which require more than two parallel corpora or large-scale monolingual corpora for both languages. Therefore, it is difficult to adopt sign language translation because even monolingual corpora for sign language are difficult to create. Our method is also based on transfer learning but requires only one parallel corpus and one large-scale monolingual corpus, so it is rather easy to be created.

Edunov et al. (2018) showed the effectiveness of back-translation to data augmentation for NMT, and Xia et al. (2019) used back-translation-based pivoting for data augmentation. Data augmentation is a mainstream technique for low-resource language translation, but we did not use it in this study because we wanted to confirm the effectiveness of a pre-trained model for translation.

Imamura and Sumita (2019) used a pre-trained model as the encoder of Transformer-based NMT. Sennrich and Zhang (2019) showed that NMT can outperform SMT for a small amount of training data using several recent techniques that have shown to be helpful in low-resource settings.

3. Our Corpus

3.1. Corpus Overview

We have been building a Japanese-Japanese Sign Language (JSL) news corpus to study Japanese to JSL machine translation. The corpus was created from daily NHK sign language news programs, which are broadcast on NHK TV with Japanese narration and JSL signing.

Feature	Description	Freq.
Nodding	Used as punctuations, topicalization, and conjunctions.	4.91
Pointing	Typically used as pronouns, but also used as emphasizing the meanings and indicating the former word as subject of the sentence.	1.75
Classifier	Morphological system that can express events and states using many morpheme.	0.26

Table 1: Special features of JSL transcribed in the corpus. Freq. represents frequency in the corpus (number of features per sentence).

JP	東京は夜から雪や雨の降る所がある見込みです。
EN	Tokyo will have places where snow and rain will fall from tonight.
SL	*Nodding*, TOKYO, R: TOKYO + L:*Pointing*, *Nodding*, DARK, FROM, *Nodding*, SNOW, *Nodding*, RAIN, *Nodding*, REGION, EXIST, DREAM, *Nodding*
JP	サッカー日本代表の新しい監督が決まりました。
EN	The new coach of the Japanese national football team has been decided.
SL	*Nodding*, SOCCER, JAPAN, REPRESENTING, NEW, GUIDANCE, WHO, DECIDE, FINISH, *Nodding*

Table 2: Examples from our corpus. JP means Japanese transcription, EN means translation of JP into English, and SL means sign language word sequences, with word segmentation of ",".

The corpus consists of Japanese transcriptions, JSL videos, and JSL transcriptions. Japanese transcriptions were transcribed by revising the speech recognition results of news programs. JSL transcriptions are carried out by changing the sign motions of the newscasters into sign word glosses. The JSL videos were manually extracted from the programs by referring to the time intervals of the transcribed JSL transcriptions. The corpus currently includes about 130,000 sentence pairs taken from broadcasts running from April 2009. In this corpus, sign languages are presented by 18 casters (11 deaf casters and 7 hearing-able interpreters).

Note that, Japanese and JSL phrase pairs are not literal translations, so there are many subject complements, omissions, and so on.

3.2. Sign Words Transcription Rules

JSL transcriptions of the corpus were manually transcribed by native JSL speakers. The words in the transcriptions are represented using the Japanese words that have the most similar meanings. We also transcribed the special features listed in Table 1, which are frequently used in JSL.

This notation method is called "glosses" in sign language research. Examples from our corpus are shown in Table 2. Note that, our transcription also includes multi-linear expressions, such as place name using the right hand and pointing with the left hand at the same time. For example in Table 1, "R:TOKYO + L:*Pointing*" means the place

name “Tokyo” is expressed with the right hand, and *Pointing* with left hand at the same time. We use only sign word sequences expressed using the right hand in this paper.

4. Translation with Pre-trained Model

As we mentioned in Section 3.1., we have only 130,000 sentence pairs in our corpus. This is far smaller than open corpora used in machine translation such as the WMT 2014 English–German dataset, which contains around 4.5M sentence pairs. Generally, sign languages do not have writing systems, so transcriptions of sign language are difficult to gather.

To overcome the shortage of training data, we use a pre-trained model as the initial model of the encoder of the translation model. An overview of our method is illustrated in Figure 1. Our method is based on Transformer (Vaswani et al., 2017) and uses a pre-trained BERT model (Devlin et al., 2019) as the initial model of the encoder. Input sentences written in spoken language are embedded using the encoder, then the embedded vectors are fed into the decoder and translated into sign language glosses. The learning process involves fine-tuning the pre-trained model and learning the decoder in parallel.

The pre-trained model can embed input Japanese sentences more relevantly than that learned from a parallel corpus, so it can help improve overall translation quality. Moreover, most of the “loss” calculated in the training process can be used to optimize the decoder due to the difference in the training rate between the encoder and decoder, so training the decoder can progress rapidly. We call our method “NMT-BERT.” Our translation model is almost the same as that Imamura and Sumita (2019) used. Our study differs in that we applied the model to sign language.

There are many techniques to improve translation models such as tied embedding, label smoothing, and data augmentation. However, we did not use them because we wanted to confirm the effectiveness of the pre-trained model in translation.

5. Experiment

5.1. Experimental Settings

Our experiments were based on our corpus mentioned in Section 3. We randomly divided the corpus into 130,215 sentence pairs for training, 1,000 pairs for development, and 2,000 pairs for testing. We also prepared reduced training datasets containing 50,000, 10,000, and 1,000 sentence pairs for comparing performance in low-data settings. We denote the 130,215 sentence pairs of training data as 130K, that of 50,000 as 50K, 10,000 as 10K, and 1,000 as 1K.

For the encoder of our method, we used our in-house Japanese BERT model learned from about 7.1 GB of Japanese Wikipedia, Twitter, News articles, and other corpora. Hyperparameters were the same as BERT-base[1], which has a 12-layer, 768 hidden states Transformer model with 12-head attention. We used SentencePiece (Kudo and Richardson, 2018) as the tokenizer for Japanese with a vocabulary size of 32,000.

[1] https://github.com/google-research/bert

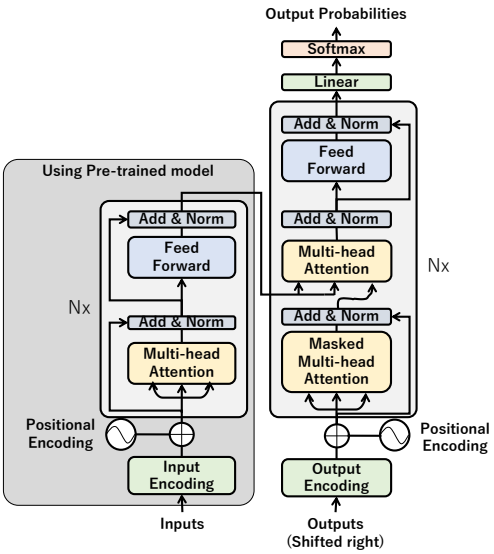


Figure 1: Overview of our method. Our method is based on Transformer and uses pre-trained model as initial model of encoder.

For the decoder, we used Transformer, which has 768 hidden states, with 8-head attention with layer normalization for each layer, and the number of layers are four for all training data, which are tuned based on the BLEU score (Papineni et al., 2002) on the development data. We used beam search in translating with a beam size of 10. We used each JSL word as a token, and words used less than 5 times in the corpus were regarded as out-of-vocabulary words (OOVs). As a result, the decoder has a vocabulary size of 5,984.

The models were implemented using pytorch[2] with Transformers[3] and learned with the Adam optimizer (Kingma and Ba, 2015) on the basis of cross-entropy loss. We used the stochastic gradient descent with warm restarts (SGDR) scheduler (Loshchilov and Hutter, 2017) without restart to adjust the learning rate with 5 epochs for warmup. The learning rates were 1.0×10^{-3} for training the decoder and 2.0×10^{-5} for fine-tuning the pre-trained model.

Other hyperparameters used were: a minibatch size of 50, dropout rate of 0.1, and 50 training iterations with early stopping on the basis of the BLEU score for the development data.

5.2. Baseline Methods

5.2.1. PBSMT Baseline

We prepared the phrase-based statistic machine translation (PBSMT) baseline method. We used Moses v4 (Koehn et al., 2007) to train for this baseline. We used MGIZA (Gao and Vogel, 2008) for word alignment and lmplz of KenLM (Heafield et al., 2013) for 5-gram language model training. We also used batch MIRA (Cherry and Foster, 2012) to optimize feature weights on the development data with the target metric of the BLEU score. We denote this method as ”PBSMT.”.

[2] https://pytorch.org/

[3] https://github.com/huggingface/transformers

Method	BLEU			
	130K	50K	10K	1K
PBSMT	23.96	22.57	19.28	12.43
NMT-Base	23.10	19.91	7.74	2.00
NMT-BERT	**24.24**	22.37	15.83	5.55

Table 3: Experimental results.

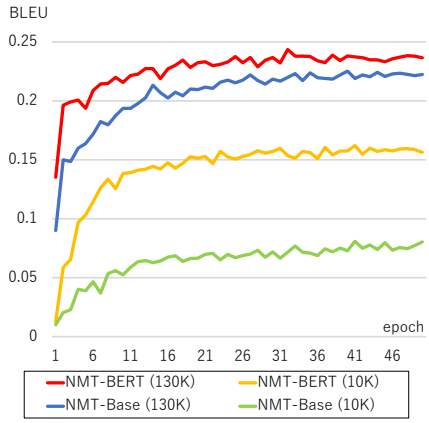


Figure 2: BLEU score for each epoch.

5.2.2. Transformer without Pre-trained Model

We used Transformer without a pre-trained model as another baseline. With this method, all parameters are trained from training data. The number of layers for the encoder and decoder were two for all training data, and other hyperparameters were the same as NMT-BERT, which were the best settings on the development data. We also used SentencePiece as a tokenizer with the model learned from the training corpus with a vocabulary size of 8,000. We denote this method as "NMT-Base."

5.3. Results

Table 3 presents the experimental results. NMT-BERT outperformed the baseline methods for 130K and outperformed NMT-Base for all datasets. Therefore, we confirmed that using a pre-trained model for NMT is effective especially in small-training-data situations. However, PBSMT is still the best for smaller datasets. NMT-BERT outperformed NMT-Base, especially for low-training-data situations. Generally, learning an NMT models requires large-scale parallel corpora. However, NMT-BERT requires only a small parallel corpus and large-scale monolingual corpus of spoken language, which are rather easy to create. This is very advantageous, especially for sign language translation, because corpora of sign language are difficult to create.

Figure 2 shows the BLEU scores for the development data for each epoch. We show two cases for the training datasets, 130K and 10K. NMT-BERT was far better in early learning processes (around epochs 1–10). The encoder learned only from the training data outputting almost random vectors in the early epochs, but the pre-trained model could output relevant vectors for the input sentence. The decoder of NMT-BERT can use these relevant vectors, so optimizing the decoder can be easier than that of NMT-Base. Moreover, the output vector of the pre-trained model represents word-to-word relations such as synonym and paraphrases, so NMT-BERT can translate OOVs or less frequent words in the training corpus using these relations. This is why NMT-BERT outperformed NMT-Base.

On the other hand, PBSMT was best for 10K and 1K. The pre-trained model is useful for improving translation quality, but there is a limit. For these very small training data situations, other techniques such as that used by Sennrich and Zhang (2019) should be used.

Excluded word	BLEU
None	24.24
Nodding	22.45
Pointing	25.19

Table 4: Results of word exclusion test.

Sign languages have special features such as *Nodding* and *Pointing*. We analyzed our translation results to investigate the effect of these special features. Table 4 shows the BLEU score of excluding *Nodding* or *Pointing* from both translation results and reference data for NMT-BERT[4]. The fact that excluding *Pointing* increases the BLEU score by around 1.0 suggests that translating *Pointing* is difficult. *Pointing* is typically used as pronouns but sometimes used to emphasize the meanings of nouns or indicate the word as the subject of the sentence, so spoken languages do not have the same word/function. This is why *Pointing* is difficult to translate. On the other hand, excluding *Nodding* lowers the BLEU score. *Nodding* is mostly used as punctuations, topicalization, and conjunctions. These functions are also used in spoken language, so *Nodding* can be translated easily.

5.4. Toward Improving *Pointing* Translation

To improve *Pointing* translation, we evaluated three translation methods. One involves translating *Pointing* as a sign word, i.e., the same as with NMT-BERT, and is denoted as Translating (Figure 3-(a)). The second method combines *Pointing* and the former word into one token and is denoted as Jointed-*Pointing* (Figure 3-(b)). If the meanings of *Pointing* are decided only by the former word, combining *Pointing* and the former word can clarify their meanings, so it may help improve translation quality. The third method involves using sequential labeling for *Pointing* and is denoted as Sequential labeling (Figure 3-(c)). Sequential labeling is typically used for finding specific parts from a sequence such as named entity recognition or part-of-speech tagging by taking into account context and grammatical rules (Ma and Hovy, 2016). We used this method to find the specific part —to use *Pointing*— from the sentence. If the decision to use *Pointing* is made by context and grammatical rules rather than the former word, Sequential labeling will work well. With Sequential labeling, we use multi-task learning for two tasks —translating into sign language and judging whether *pointing* is needed for the translated word—.

[4]We did not analyze for *Classifier*, which is a special features of sign language. This is because *Classifier* plays an important role for the meanings of a sentence, so excluding *classifier* make a sentence meaningless, so the evaluation would not make sense.

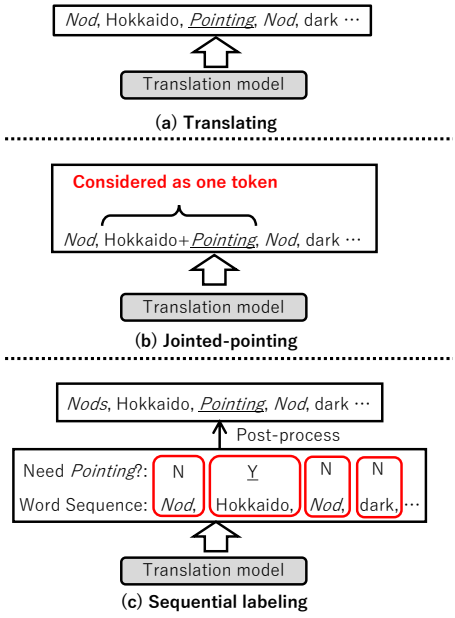


Figure 3: Three translation methods specially designed for translating *Pointing*.

Method	BLEU
Translating	**24.24**
Jointed-pointing	23.39
Sequential labeling	22.12

Table 5: Results of translation methods specially designed for *Pointing* translation.

We used our corpus (130K) to evaluate these methods, and the results are listed in Table 5. The BLEU score was the best for Translating. This suggest that the decision to use *Pointing* is made by neither only the former word, only the context nor only grammatical rules. We believe it is decided from the context of sign word sequences, so translating *Pointing* should be done by considering the context of sign word sequence, as language models do. However, there is still room for improving the translation accuracy for *Pointing*. This is left as our future work.

6. Conclusion and Future Work

In this paper, we presented a neural machine translation method from Japanese to Japanese Sign Language glosses using a pre-trained model as the initial model of the encoder, and confirmed that the method works well, especially in small-training-data situations. The BLEU scores for the method was 24.24 using training data of about 130,000 sentence pairs, which outperformed the baseline methods including phrase based statistical machine translation, which had a BLEU score of 23.96. Using a pre-trained model is better than learning models only from training data, especially in small-training-data situations.

We also conducted trials to improve the translation quality for *Pointing*. The results indicate that *Pointing*, which is a special feature of sign language, should be translated considering long-term dependencies.

We showed that Transformer with a pre-trained model can be used with a small amount of training data, so we can apply many techniques designed for use with Transformer such as tied embedding, label smoothing, and data augmentation. Using these techniques is for our future work. To improve the translation quality of special features of sign language such as *Pointing* is also for future work.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments.

7. Bibliographical References

Azuma, M., Hiruma, N., Sumiyoshi, H., Uchida, T., Miyazaki, T., Umeda, S., Kato, N., and Yamanouchi, Y. (2018). Development and evaluation of system for automatically generating sign-language CG animation using meteorological information. In *International Conference on Computers Helping People with Special Needs*, pages 233–238. Springer.

Cherry, C. and Foster, G. (2012). Batch tuning strategies for statistical machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 427–436. Association for Computational Linguistics.

Cihan Camgoz, N., Hadfield, S., Koller, O., Ney, H., and Bowden, R. (2018). Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7784–7793.

Cormier, K., Schembri, A., and Woll, B. (2013). Pronouns and pointing in sign languages. *Lingua*, 137:230–247.

Dabre, R., Fujita, A., and Chu, C. (2019). Exploiting multilingualism through multistage fine-tuning for low-resource neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1410–1416, Hong Kong, China, November. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Edunov, S., Ott, M., Auli, M., and Grangier, D. (2018). Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500.

Gao, Q. and Vogel, S. (2008). Parallel implementations of word alignment tool. In *Software engineering, testing, and quality assurance for natural language processing*, pages 49–57.

Hanke, T. (2004). HamNoSys-representing sign language data in language resources and language processing contexts. In *LREC*, volume 4, pages 1–6.

Heafield, K., Pouzyrevsky, I., Clark, J. H., and Koehn, P. (2013). Scalable modified Kneser-Ney language model

estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696.

Imamura, K. and Sumita, E. (2019). Recycling a pre-trained BERT encoder for neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 23–31.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations 2015*.

Kipp, M., Heloir, A., and Nguyen, Q. (2011). Sign language avatars: Animation and comprehensibility. In *International Workshop on Intelligent Virtual Agents*, pages 113–126. Springer.

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. *ACL 2017*, page 28.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 66–71.

Lample, G., Ott, M., Conneau, A., Denoyer, L., and Ranzato, M. (2018). Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049.

Loshchilov, I. and Hutter, F. (2017). SGDR: Stochastic gradient descent with warm restarts. In *IInternational Conference on Learning Representations 2017*.

Ma, X. and Hovy, E. (2016). End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1064–1074.

Massó, G. and Badia, T. (2010). Dealing with sign language morphemes in statistical machine translation. In *4th workshop on the representation and processing of sign languages: corpora and sign language technologies, Valletta, Malta*, pages 154–157.

Ministry of Internal Affairs and Communications. (2019). Achievements of closed caption etc. for TV programs in 2018. Press release (in japanese).

Mocialov, B., Hastie, H., and Turner, G. (2018). Transfer learning for British Sign Language modelling. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 101–110, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Morrissey, S. (2011). Assessing three representation methods for sign language machine translation and evaluation. In *Proceedings of the 15th annual meeting of the European Association for Machine Translation (EAMT 2011), Leuven, Belgium*, pages 137–144.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Romeo, M., Evans, A., Pacheco, D., and Blat, J. (2014). Domain specific sign language animation for virtual characters. In *2014 International Conference on Computer Graphics Theory and Applications (GRAPP)*, pages 1–8. IEEE.

San-Segundo, R., Montero, J. M., Córdoba, R., Sama, V., Fernández, F., D ' Haro, L., López-Ludeña, V., Sánchez, D., and García, A. (2012). Design, development and field evaluation of a Spanish into sign language translation system. *Pattern Analysis and Applications*, 15(2):203–224.

Sennrich, R. and Zhang, B. (2019). Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy, July. Association for Computational Linguistics.

Skorokhodov, I., Rykachevskiy, A., Emelyanenko, D., Slotin, S., and Ponkratov, A. (2018). Semi-supervised neural machine translation with language models. In *Proceedings of the AMTA 2018 Workshop on Technologies for MT of Low Resource Languages (LoResMT 2018)*, pages 37–44, Boston, MA, March. Association for Machine Translation in the Americas.

Stein, D., Schmidt, C., and Ney, H. (2010). Sign language machine translation overkill. In *International Workshop on Spoken Language Translation (IWSLT) 2010*.

Stoll, S., Camgöz, N. C., Hadfield, S., and Bowden, R. (2018). Sign language production using neural machine translation and generative adversarial networks. In *Proceedings of the 29th British Machine Vision Conference (BMVC 2018)*. British Machine Vision Association.

Uchida, T., Sumiyoshi, H., Miyazaki, T., Azuma, M., Umeda, S., Kato, N., Yamanouchi, Y., and Hiruma, N. (2018). Evaluation of a sign language support system for viewing sports programs. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 361–363.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Xia, M., Kong, X., Anastasopoulos, A., and Neubig, G. (2019). Generalized data augmentation for low-resource translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5786–5796, Florence, Italy, July. Association for Computational Linguistics.

Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 145–150
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

Towards Large-Scale Data Mining for Data-Driven Analysis of Sign Languages

Boris Mocialov[1], **Graham Turner**[2], **Helen Hastie**[3]
[1]School of Engineering and Physical Sciences, Heriot-Watt University, Edinburgh, UK
[2]School of Social Sciences, Heriot-Watt University, Edinburgh, UK
[3]School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh, UK
{bm4, g.h.turner, h.hastie}@hw.ac.uk

Abstract

Access to sign language data is far from adequate. We show that it is possible to collect the data from social networking services such as TikTok, Instagram, and YouTube by applying data filtering to enforce quality standards and by discovering patterns in the filtered data, making it easier to analyse and model. Using our data collection pipeline, we collect and examine the interpretation of songs in both the American Sign Language (ASL) and the Brazilian Sign Language (Libras). We explore their differences and similarities by looking at the co-dependence of the orientation and location phonological parameters.

Keywords: Sign Language Data Mining, Phonological Parameter Co-Dependence

1. Introduction

The data-driven field of automated sign language understanding is dependent on large amounts of high-quality data, independent of the application or the motivation of the research. Unfortunately, such data have typically had restricted access, either due to the projects finishing or limiting license terms. Online services, on the other hand, offer large amounts of data that have more relaxed terms and conditions and are available for as long as the service providers exist. The field of written and spoken languages has recently greatly benefited from the use of such data for natural language understanding projects trained on, for example, Reddit, Twitter, Amazon reviews.

However, there is an uneven distribution of users of spoken languages versus sign language users around the world. Therefore, the amount of sign language data on the internet is naturally lower than that of spoken languages. Moreover, sign languages do not have a common writing system as opposed to spoken languages, which makes it very difficult to annotate. Furthermore, the data from research projects usually have application-dependent annotation using either words in written languages, phonological parameters, or sign pictures (Konrad, 2015). Despite the fact that there is no common writing system for sign languages, HamNoSys defines one notation system that is often used by researchers. This system distinguishes phonological parameters (e.g. location, orientation, movement, handshape, non-manual gestures) present during signing (Hanke, 2004).

First, we show that it is possible to utilise social networking platforms to support research in data-driven automated sign language understanding. Second, we take a look at two sign languages that have relatively little historical relationship. One sign language being ASL and the second being Libras and investigate the signing behaviour of the spoken song interpreters, while looking at three English songs: 'Love Yourself' by Justin Bieber, 'Halo' by Beyoncé, and 'Love On The Brain' by Rihanna. We investigate and compare two phonological parameters: hand location relative to the signers' body and extended finger orientation. This work will quantify frequently occurring hand positioning during the signing and compare the prevailing hand positions and orientations between the two sign languages, aiming to show that sign languages evolve differently. The reason why we investigate interpreted songs is because we want to compare sign languages by looking at continuous signing in different sign languages that sign the same information. Findings in this paper could also assist researchers who work on developing models for sign language understanding by reducing the search space of the models during the optimisation by ignoring combinations that are relatively infrequent during continuous signing.

2. Motivation for Automated Processing of Sign Languages

With the rise of accurate pose prediction and hand estimation libraries such as the OpenPose (Cao et al., 2018; Simon et al., 2017; Cao et al., 2017; Wei et al., 2016), researchers in the field of automated sign language understanding are now able to focus on high level abstract research ideas. Contemporary research looks at translating written languages to sign languages and vice versa, thus resembling research done in the field of the machine translation for spoken languages (Camgöz et al., 2018; Stoll et al., 2019; Yuan et al., 2019).

The common linguistically inspired approach is for the raw visual modality of the sign languages to be broken down into the sub-lexical phonological parameters (e.g. location, orientation, movement, handshape, and non-manual gestures). Multiple previous works have modelled individual phonological parameters. Cooper and Bowden (2007) modelled hand location, movement, and relative hand position and called them the sub-sign units. Cooper et al. (2012) relied on handshape, location, movement, and relative hand position in their work on recognition of the individual signs in The British Sign Language. Buehler et al. (2009) examined movement, handshape, and orientation while matching the combination of these parameters to find similar signs. Also Buehler et al. (2010) used location and handshape in the multiple instance learning problem. Koller et al. (2016)

focused on modelling sixty handshapes. Their model is a chain of convolutional neural networks (VGG) pre-trained on the ImageNet data (Simonyan and Zisserman, 2015). In our work, we generate linguistic annotations in the form of hand location relative to the signers' body and extended finger orientation for the continuous interpretations of the three English songs (mentioned above).

The lack of the text annotation that could provide context remains an issue for video data. Joze and Koller (2018) noticed that many signing videos have captions, which could be an additional source of annotation, as more and more content is being generated online, including that for the deaf community. In this work, we focus only on the linguistic annotations without inferring the context.

Type	Motivation
Recognition	Context-Specific (Ko et al., 2019) Isolated Signs (Zhou et al., 2009) Individual Phonological Parameters (Cooper et al., 2012)
Translation	Sign-Text (Camgöz et al., 2018) Text-Sign (Stoll et al., 2019)
Learning	Zero-Shot (Bilge et al., 2019) Clustering (Nandy et al., 2010) Augmentation (Mocialov et al., 2017)
Linguistic Studies	Phonological Parameter Co-Dependence (Östling et al., 2018)
Education	Teaching (Stefanov and Beskow, 2017) Edutainment (Zafrulla et al., 2011)
Sign Spotting	Queries (Belissen, 2018)

Table 1: Research directions in the field of the automated sign language understanding

We group surveyed papers by their motivation, omitting works that use data other than a single RGB camera and papers that focus on pose estimation, tracking, or finger spelling, as these do not directly align with research in sign language understanding. Table 1 categorises the research directions in the field of automated sign language understanding. It can be seen that there are projects that focus on more abstract concepts than learning the isolated signs, such as automated data-driven sign language translation. Apart from the recognition of the signs as an attempt to bridge the gap between the hearing and the deaf communities, assistive tools for digital sign language content annotation are gaining interest. For example, Takayama and Takahashi (2018) automatically annotate datasets and Belissen (2018) query databases with videos of signs, which could be beneficial for accelerating research in linguistic aspects of sign language.

3. Methodology

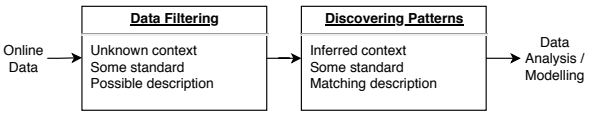


Figure 1: Data collection pipeline for collecting the signing data from the social networking services for the purpose of data-driven automated sign language understanding

Data from online social-media resources tends to be very unpredictable. Therefore, the collection of data has to pass through a number of stages as we suggest in Figure 1. The first stage of the data collection pipeline is data filtering, where we turn the online data that has no standard into data that has some pre-defined standard. The second stage looks for the patterns in the filtered data either with the help of metadata or the automatic visual analysis of the collected filtered data.

3.1. Online Data

In this research, we focus on interpreted songs. As deaf-specific music performers are rare, sign language users resort to 'listening' to interpretations of the songs found in the spoken or written languages that are being interpreted by those who can both hear and sign. This is evidenced by the relatively large amount of content found in online resources such as TikTok, YouTube, or Instagram. Such content makes the interpretation of the spoken songs possible for the deaf community, encouraging visualisation of music (Desblache, 2019). We consider interpreted songs as our data format because it is possible to find the same songs interpreted in different sign languages, which makes the comparison of the sign languages more precise. We collect one video for every interpreted song for each sign language. Therefore, we have collected a small dataset of continuous signing videos from YouTube from six different signers, interpreting three songs in two sign languages for this proof of concept study.

3.2. Data Filtering

We use the OpenPose library for data filtering. The library detects 2D or 3D anatomical key-points, associated with the human body, hands, face, and feet in a single image. The library provides 21 (x,y) key-points for every part of the hand, 25 key-points for the whole body skeleton, and 70 key-points for the face. The OpenPose library helps us apply simple filters to the raw data, discarding all the content that has more than one signer at the same time or any heavy obstructions or occlusions. We also discard the content that has too few key-points visible, as we think it is essential to see the upper body and the hands to make sense of the signing. By performing such filtering, we enforce a quality standard upon the collected online data. However, the context and the signer profile remain unknown.
Other filters could include normalisation, transformation, and rotation of key-points to make the signer appear the same size across videos and to make signers face the camera for more accurate modelling.

3.3. Discovering Patterns

We perform visual analysis by extracting the location and orientation sub-lexical phonological parameters from the filtered data looking at the frequency of occurrences of specific location/orientation combinations in the collected filtered data. The following sections will show how we infer sub-lexical components by making use of identified key-points and geometry.

Likewise, metadata can assist in discovering patterns. This metadata can comprise of hashtags, textual description, or the embedded captions on the videos.

3.3.1. Extended Finger Orientation

A total of eight orientations have been used for the extended finger orientation as defined in the HamNoSys notation with each orientation having $45°$ movement (north, north-east, east, south-east, south, south-west, west, and north-west). HamNoSys does define more orientations (e.g. towards or away from the body), however having 2D data makes it difficult to estimate additional orientations.

The angle is calculated using the inverse trigonometric function between the radius and middle finger coordinates as follows:

$$-\pi/2 < \arctan(q_y - p_y, q_x - p_x) < \pi/2,$$

where q and p are the (x, y) coordinates of radius and middle finger metacarpal bones with every orientation having $\pi/4$ freedom

3.3.2. Hand Location Relative to the Body

A total of six locations around the body have been used to determine hand position (ears, eyes, nose, neck, shoulder, and abdomen), as opposed to the forty six defined by the HamNoSys notation system. Six were chosen to simplify the detection while complying with the OpenPose library standards. Hand centroids are calculated as follows:

$$centroid_{right} = (\sum_{i=1}^{N} x_{i_{right}}/N, \sum_{i=1}^{N} y_{i_{right}}/N)$$

$$centroid_{left} = (\sum_{i=1}^{N} x_{i_{left}}/N, \sum_{i=1}^{N} y_{i_{left}}/N)$$

where N is the number of points provided by the OpenPose library for each hand.

In order to assign the relative hand location, a threshold has to be assigned as to how far the centroid of a hand can be from a specific body location so as to still be relatively close to that body part. All the distances are measured in pixels and the threshold is set to be 10% of the diagonal of the image frame, which is approximately 100 pixels. If the distance of a centroid away from all the body parts exceeds the threshold, the hand is considered to be in the 'neutral signing space'.

The distance matrix D for every hand is calculated as follows:

$$M_r \dots N_r = |q_{m\dots n} - centroid_{right}|$$

$$M_l \dots N_l = |q_{m\dots n} - centroid_{left}|$$

$$D = \begin{pmatrix} M_r & M_l \\ \vdots & \vdots \\ N_r & N_l \end{pmatrix}$$

Where $q_{m\dots n}$ are the (x, y) position of the body parts, defined by the OpenPose library (e.g. nose, neck, shoulder, elbow, etc.) and the $M_r \dots N_r$ and $M_l \dots N_l$ are the Euclidean distances between the body parts and right and left hand centroids.

In order to find the body part B_{right} or B_{left}, which has the smallest distance to the centroid of the right or left hand, we use

$$B_{right} = \operatorname{argmax} D_{i,1}$$

$$B_{left} = \operatorname{argmax} D_{i,2}$$

The distances are then compared to a threshold to determine if a hand is near a particular body part or is in the 'neutral signing space' anywhere around the body.

3.4. Data Analysis / Modelling

Once the data has been filtered and the patterns have been discovered, we acquired information on 43016 hand locations and the same number for the hand orientations for ASL and 38258 for both hand location and orientation for Libras for the interpreted three songs. We are interested in the analysis of the co-dependence of phonological parameters for each hand and comparing the significant co-dependences across the two sign languages.

3.4.1. Phonological Parameter Co-Dependence

Here, we refer to location as TAB and to orientation as ORI for the shorthand notation. First, a global C_{ORI_N, TAB_M} contingency table is generated and counts the occurences for both location/orientation variables for every category that occurs in the collected data (e.g. North, North-East, etc. for orientation and Shoulder, Neck, etc. for location). Second, a series of local contingency tables $C_{2\times2}$ are constructed from the global C_{ORI_N, TAB_M} contingency table for every category of every variable as a post-hoc step. Finally, Bonferroni-adjusted p-value was used (Bland and Altman, 1995) to check if the presence of a particular location/orientation combination in the data set is significant, compared to other location/orientation combinations, by performing a Chi-square test of independence of variables for all the $C_{2\times2}$ contingency tables.

$$C_{ORI_N, TAB_M} = \begin{pmatrix} c_{ORI_1, TAB_1} & \dots & c_{ORI_1, TAB_M} \\ \vdots & \ddots & \vdots \\ c_{ORI_N, TAB_1} & & c_{ORI_N, TAB_M} \end{pmatrix}$$

$$S_{ORI_N, TAB_M} = \sum C_{ORI_N, TAB_M} \cap (C_{ORI_i, TAB_j} \cup \sum ORI_i \cup \sum TAB_j)$$

$$C_{2\times2} = \begin{pmatrix} C_{ORI_i, TAB_j} & \sum ORI_i \\ \sum TAB_j & S_{ORI_N, TAB_M} \end{pmatrix}$$

4. Preliminary Results

4.1. Online Data

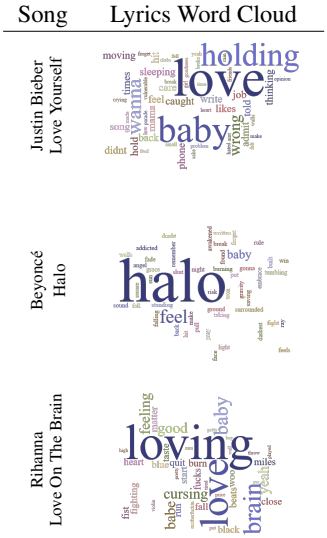


Table 2: Word cloud generated from the lyrics for the three English language songs

Table 2 shows the word cloud for the three songs ('Love Yourself' by Justin Bieber, 'Halo' by Beyoncé, and 'Love On The Brain' by Rihanna) generated from the lyrics obtained online. The purpose of the word cloud is to give insight into which words are frequent in the lyrics. As it can be seen, the lyrics for all the songs often mention love and romantic feelings.

Table 3: Screenshots of the collected online data for three songs by three different artists with English spoken language interpreted by six different signers, three signers per sign language

Table 3 shows screenshots of the collected online data for three English songs performed by three different artists. The songs are interpreted in two sign languages by different signers. From the screenshots, it can be seen that the proximity of the signer to the camera varies. Some videos are edited by applying black and white or vintage camera filters. As a general rule, there is no camera movement, but the signers usually dance slightly to the songs.

4.2. Data Filtering

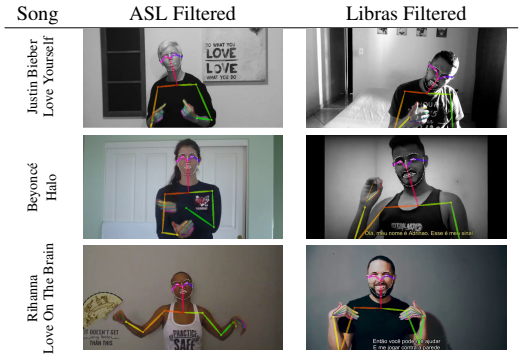


Table 4: Screenshots of the filtered online data after applying the OpenPose library. Visible skeletons on the screenshot means that the library was able to detect a human in the video and is tracking the pose, hand, and face key-points

Table 4 shows screenshots of the filtered data after the OpenPose library has been applied to the data. The library was able to detect a human in the video and is tracking the pose, hand, and face key-points. Since we are not interested in keeping the integrity of the sequences of the frames, we simply discard the frames where key-points were not detected by the library.

4.3. Discovering Patterns between Sign Languages

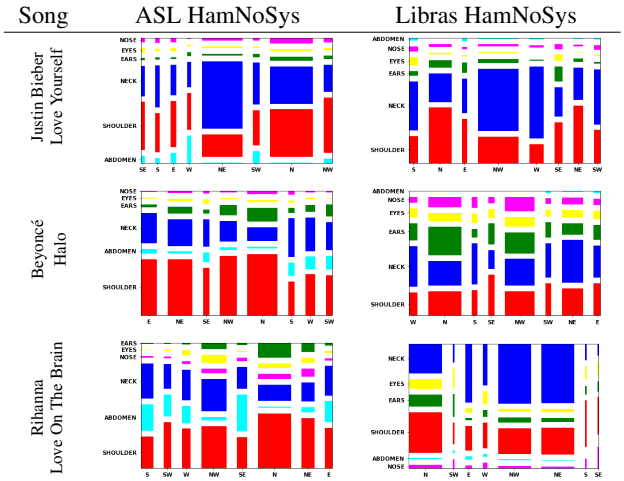
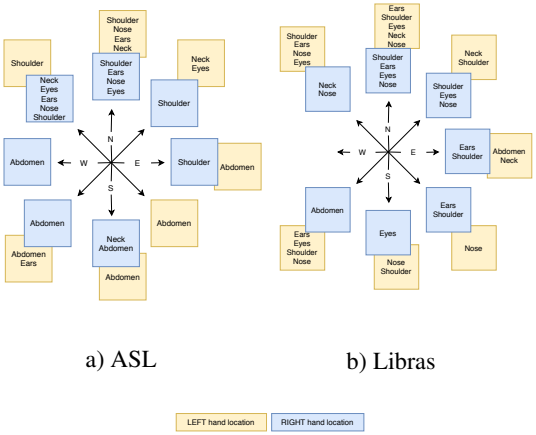


Table 5: Location/orientation relative frequencies for each video for each sign language

Table 5 shows the relative frequencies of the location/orientation combinations for each video and each sign language. We can observe that the Libras, on one hand, has less abdomen activity than the ASL (indicated in light blue) while, on the other hand, Libras has more neck and ears activity than the ASL (dark blue and green respectively). Both sign languages have more pointing up direction of the hands as opposed to other possible directions (wider NE/N/NW columns).

4.4. Data Analysis / Modelling



a) ASL b) Libras

Figure 2: Significant location and orientation phonological parameter co-dependences in a) ASL and b) Libras with Bonferroni-adjusted Chi-squared p-value < 0.001 for both sign languages

Having visually analysed the frequencies of the location/orientation combinations, we are interested in finding the significant combinations for each hand that prevails in the collected data and compare the two sign languages based on this analysis.

4.4.1. Phonological Parameter Co-Dependence

Figure 2 shows significant location/orientation co-dependences for each sign language after the Bonferroni-adjusted Chi-squared p-value analysis. We can see that both hands tend to point up at the upper side of the body, which is similar for the both sign languages. Libras, however, has more activity with both hands at the upper part of the body than the ASL. As a matter of fact, Libras has more activity with both hands around all the parts of the body. In ASL, on the other hand, the left hand is less mobile than the right hand. This could be explained by the fact that the signers in Libras were left-handed, but we do not have this information available to verify this speculation. Some significant co-dependences are unusual, for example, pointing down at the upper body level, which may feel unnatural and slightly contradicts the past findings by Cooper et al. (2012) stating that a subset of the 'comfortable' hand configurations are assumed more often during the signing, independent of the sign language. This can also be explained by the fact that the signers in the video are slightly dancing to the music, which may affect the signing orientation.

It is worth mentioning that the co-dependence analysis results of the two languages may change with the data. For example, if songs with a different sentiment were taken for the analysis. More data is needed to experiment this further.

5. Conclusion

In this work, we have showed the preliminary results of mining sign language data acquired from the internet for automated data-driven sign language processing. We have created a pipeline that downloads the videos of the interpreted songs from the internet, applies filtering of the data and then finds patterns in the data based on the HamNoSys notation that is often used for the annotation of the sign languages. This method could also be used for querying videos in large datasets. Finally, we compare two historically different sign languages (ASL and Libras) by their location/orientation co-dependencies present in the collected data and show that, despite there being little historical background of the two languages interacting, they still share similar signing patterns with small variations in the flexibility of the hands, which can be explained by the fact that people converge to the usage of the 'comfortable' hand configurations.

Future work will compare even more historically unrelated sign languages and look at the interpretations of a greater number of songs, in order to have a more accurate comparison of the signing patterns across the sign languages.

6. Acknowledgements

This work was supported by the Heriot-Watt University School of Engineering & Physical Sciences James Watt Scholarship and Engineering and Physical Sciences Research Council (EPSRC), as part of the CDT in Robotics and Autonomous Systems at Heriot-Watt University and The University of Edinburgh (Grant reference EP/L016834/1)

7. Bibliographical References

Belissen, V. (2018). Sign language video analysis for automatic recognition and detection. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*.

Bilge, Y. C., Ikizler-Cinbis, N., and Cinbis, R. G. (2019). Zero-shot sign language recognition: Can textual data uncover sign languages? In *Proceedings of the British Machine Vision Conference*.

Bland, J. M. and Altman, D. G. (1995). Multiple significance tests: the bonferroni method. *BMJ*, 310(6973):170.

Buehler, P., Zisserman, A., and Everingham, M. (2009). Learning sign language by watching tv (using weakly aligned subtitles). In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*.

Buehler, P., Everingham, M., and Zisserman, A. (2010). Employing signed tv broadcasts for automated learning of British Sign Language. In *Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages*.

Camgöz, N. C., Hadfield, S., Koller, O., Ney, H., and Bowden, R. (2018). Neural sign language translation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*.

Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. (2017). Realtime multi-person 2D pose estimation using part affinity fields. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*.

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., and Sheikh, Y. (2018). OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*.

Cooper, H. and Bowden, R. (2007). Large lexicon detection of sign language. In *Proceedings of the International Workshop on Human-Computer Interaction*.

Cooper, H., Ong, E.-J., Pugeault, N., and Bowden, R. (2012). Sign language recognition using sub-units. *Journal of Machine Learning Research*, 13(Jul):2205–2231.

Desblache, L., (2019). *How is Music Translated? Mapping the Landscape of Music Translation*, pages 219–264. Palgrave Macmillan UK, London.

Hanke, T. (2004). Hamnosys-representing sign language data in language resources and language processing contexts. In *Proceedings of the Workshop on Representation and processing of sign languages (LREC 2004)*.

Joze, H. R. V. and Koller, O. (2018). MS-ASL: A large-scale data set and benchmark for understanding American Sign Language. In *Proceedings of the British Machine Vision Conference*.

Ko, S.-K., Kim, C. J., Jung, H., and Cho, C. (2019). Neural sign language translation based on human keypoint estimation. *Applied Sciences*, 9(13):2683.

Koller, O., Ney, H., and Bowden, R. (2016). Deep hand: How to train a cnn on 1 million hand images when your data is continuous and weakly labelled. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*.

Konrad, R. (2015). DGS corpus annotation guidelines. In *Proceedings of Digging into Signs Workshop: Developing Annotation Standards for Sign Language Corpora*.

Mocialov, B., Turner, G., Lohan, K. S., and Hastie, H. (2017). Towards continuous sign language recognition with deep learning. In *Proceedings of Workshop of Creating Meaning With Robot Assistants (Humanoids 2017)*.

Nandy, A., Prasad, J. S., Mondal, S., Chakraborty, P., Nandi, G. C.", e. V. V., Vijayakumar, R., Debnath, N. C., Stephen, J., Meghanathan, N., Sankaranarayanan, S., Thankachan, P. M., Gaol, F. L., and Thankachan, N. (2010). Recognition of Isolated Indian Sign Language Gesture in Real Time. In *Proceedings of the Conference on Information Processing and Management*.

Östling, R., Börstell, C., and Courtaux, S. (2018). Visual iconicity across sign languages: Large-scale automated video analysis of iconic articulators and locations. *Frontiers in psychology*, 9:725.

Simon, T., Joo, H., Matthews, I., and Sheikh, Y. (2017). Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations*.

Stefanov, K. and Beskow, J. (2017). A Real-Time Gesture Recognition System for Isolated Swedish Sign Language Signs. In *Proceedings of the 4th European and 7th Nordic Symposium on Multimodal Communication (MMSYM)*.

Stoll, S., Camgöz, N. C., Hadfield, S., and Bowden, R. (2019). Text2sign: Towards sign language production using neural machine translation and generative adversarial networks. *International Journal of Computer Vision*, pages 1–18.

Takayama, N. and Takahashi, H. (2018). Sign words annotation assistance using Japanese Sign Language words recognition. In *Proceedings of the International Conference on Cyberworlds (CW)*.

Wei, S.-E., Ramakrishna, V., Kanade, T., and Sheikh, Y. (2016). Convolutional pose machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*.

Yuan, T., Sah, S., Ananthanarayana, T., Zhang, C., Bhat, A., Gandhi, S., and Ptucha, R. (2019). Large scale sign language interpretation. In *Proceedings of the 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–5. IEEE.

Zafrulla, Z., Brashear, H., Presti, P., Hamilton, H., and Starner, T. (2011). CopyCat: An American Sign Language game for deaf children. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition*, pages 647–647, March.

Zhou, Y., Chen, X., Zhao, D., Yao, H., and Gao, W. (2009). Adaptive sign language recognition with exemplar extraction and map/ivfs. *IEEE signal processing letters*, 17(3):297–300.

Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 151–156
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

Extending a Model for Animating Adverbs of Manner in American Sign Language

Robyn Moncrief
School of Computing, DePaul University
Chicago, IL
rkelley5@mail.depaul.edu

Abstract

The goal of this work is to show that a model produced to characterize adverbs of manner can be applied to a variety of neutral animated signs to be used towards avatar sign language synthesis. This case study presents the extension of a new approach that was first presented at SLTAT 2019 in Hamburg for modeling language processes that manifest themselves as modifications to the manual channel. This work discusses additions to the model to be effective for one-handed and two-handed signs, repeating and non-repeating signs, and signs with contact.

Keywords: avatar technology, sign synthesis, adverbs in ASL

1. Introduction

A signing avatar is a necessary component of any sign translation system. It needs to produce signed utterances that are grammatically correct and easy to read. Although an avatar's appearance is important to its legibility, its motion has even greater impact (Malala et al., 2018). *Sign synthesis* is the computation of an avatar's movement. It is a combination of 1) path computation, which is related to motion planning (Barraquand & Latombe, 1991), 2) timing along that path, 3) the determination of joint participation in creating the path (McDonald, 2005; McDonald et al., 2016) and 4) ancillary motions required to support the clarity of the utterance (Schnepp et al, 2012). This case study presents the extension of a new approach to modeling sign language processes that manifest themselves as modifications to the manual channel.

2. Adverbs of Manner

American Sign Language (ASL) is an independent natural language (Valli & Lucas, 2000). Automatic translation between spoken language to sign have lagged behind spoken-to-spoken translation, due in part to the fact that there is no one-to-one mapping from ASL to English.

A case in point is the inclusion of adverbs of manner that express how the action of the verb takes place.

In English, the actions, states, and sensations of a verb can be modified through the application of an adverb of manner (Valli & Lucas, 2000). The following are two examples of adverbs of manner used in English. Each example is made up of two sentences. The first without the adverb; the second with the adverb:

The boy ran.
The boy ran quickly.

The couple danced.
The couple danced beautifully.

3. Related Work

Previous research of the use of adverbs of manner in ASL was limited. In contrast to English, ASL does not necessarily add an independent lexical item to express an adverb of manner. Instead, adverbs of manner are considered to be non-manual (Bickford, 2006) and modify the verb. Adverbs of manner occur through changes to the "quality of motion" as well as nonmanual signals (Baker & Cokely, 1980, Valli & Lucas, 2000, Padden, 2016).

Therefore, the starting point for related work is grounded in analysis of gesture motion. Two important examples of previous work in this area are the EMOTE (Expressive MOTion Engine) model (Chi et al, 2000) and GRETA (Hartmann, Mancini, and Pelachaud 2005). EMOTE stems from the Laban Movement Analysis (LMA) and the Effort-Shape model. The Effort-Shape model draws on LMA's classification of motion in the following way: Effort, qualitative descriptions of energy in motion and Shape, how the body changes forms during motion. EMOTE proposed a parameterization of Effort using qualitative descriptions of energy in motion through the following: 1) Space, 2) Weight, 3) Time, and 4) Flow.

GRETA's expressivity parameterization stems from psychology and expands on EMOTE's techniques for synthesis. It is comprised of six attributes: 1) Overall Activation, 2) Spatial Extent, 3) Temporal Extent, 4) Fluidity, 5) Power, and 6) Repetition.

Although researchers have examined the effects of affect on gesture (Kleinsmith & Bianchi-Berthouze, 2012), and Zhao et al. (2000) suggests that the EMOTE system would be useful in synthesizing sign languages, no one has reported on using such an approach for portraying adverbial modifiers in sign language.

Sign synthesis requires more specification than what is outlined in either EMOTE and GRETA. The characterizations proposed by EMOTE do not fully capture adverbial modifiers used in ASL. For example, EMOTE would characterize

slowly
WALK (1)

as *Bound: controlled or restrained*. As to be discussed in Section 8., data collected in this study demonstrates that the motion does not conform to this EMOTE descriptor. These systems do not take into account the importance of positionings of signs, as well as the animation techniques needed for building realism.

A more complete motion model is necessary to allow a 3D avatar to modify signs such as verbs, while supporting and respecting ASL's grammatical structure. Accuracy and

naturalness in the generated motion are necessary to make sentences as easy to understand as possible.

4. Procedure

Achieving an improved model required several steps, including the selection of adverbs, a motion study of the adverbs, animating and validating the exemplars, and data analysis (Moncrief, 2019).

4.1 Adverb Selection

Four commonly used adverbs of manner were chosen for this study. These adverbs represent different qualities of possible adverbial modifications to a sign, as well as having corresponding independent lexical items. The four adverbs contain two pairs of contrast: intensity and affect. For the contrast in intensity, the adverbs *quickly* and *slowly* were chosen. For the contrast in affect, the adverbs *happily* and *sadly*. All four adverbs were applied to the sign WALK, which is a two-handed noncontact sign with repeating motion that has few additional constraints to consider when applying motion modification.

4.2 Motion Study

Video recordings of a fluent signer are the basis for characterization of adverbial modifications and nonmanual signals. Based on the data from the video recordings, animations were generated. These animations then went through two revision cycles with a certified sign-language interpreter for grammaticality and naturalness.

4.3 Data Analysis

To create each animation, keyframe data was set by an animator. When the animations were generated, the in-betweens were interpolated. This interpolated motion data was then collected from the generated animations and used for analysis. This included joint positions for the wrists, elbows, and shoulders; timing; and joint velocities.

To determine the primary contributing variables for separating the adverbs of motion using the collected motion path data, a Linear Discriminant Analysis (Eisenbeis, 1972) was performed. This led to a high degree of separation of the adverbs, with the first linear discriminate accounting for over 98% of the separation of adverbs. The primary variables that accounted for the differences in the adverbs of manner included the wrist position and velocity. The separation based on the first two discriminant functions is shown in Figure 1.

Figures 2 and 3 show the distinct motions paths for the right wrist in the five animations on the transverse plane (x, y), as looking down at the signing space, and sagittal plane (y, z), as looking at the side of the signing space. The color variation through the motion path accounts for the change in velocity.

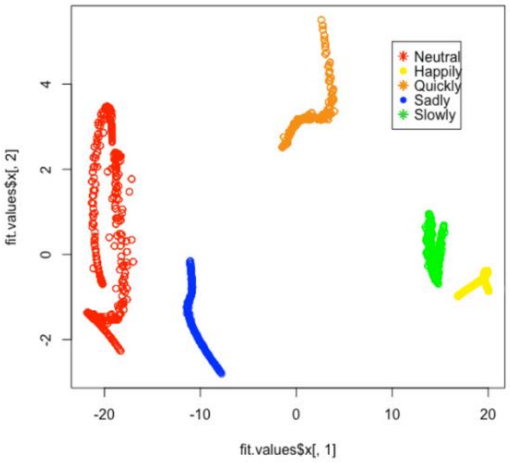


Figure 1: Separation of adverbs through the first two discriminate functions

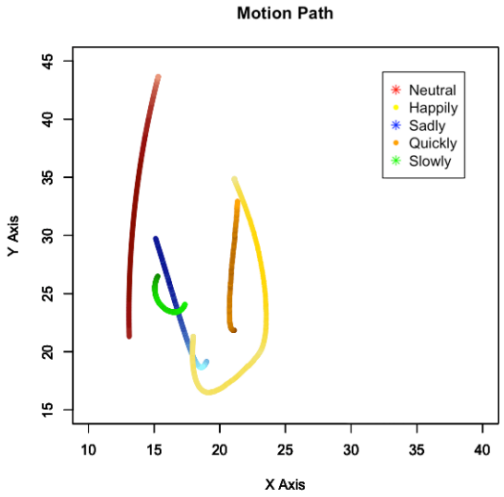


Figure 2: Motion path of wrist in the traverse plane

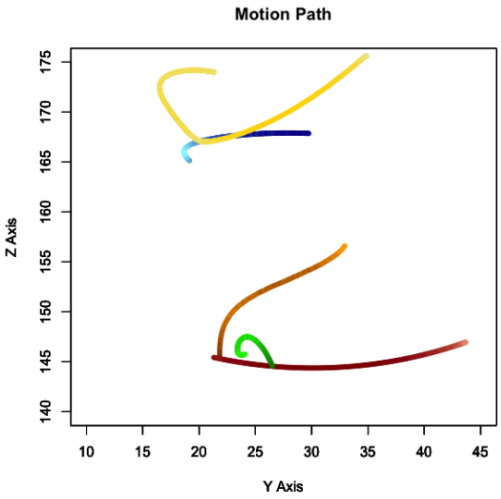


Figure 3: Motion path of wrist in the sagital plane

5. WALK Results

This model was then applied to the neutral animation of WALK for initial comparison and revision. Though wrist position and velocity were the significant contributors, those alone were not enough to convincingly convey the chosen adverbs. The model further incorporated changes to timing, ease-in and ease-out, joint positions, joint rotations, and adjustments to the spine. For the use of the *slowly* modifier, if only the wrist position and velocity were implemented, the sign would look Bound according to EMOTE parametric characterization, but based on the data and the visual observations, *slowly* required expressive joint rotations that could be described as Light.

The adjustments to the timing, joint rotations, and joint positions of the neutral animation consisted of developing a multiplier for each adverb. To account for the differences in speed along the motion path shown for each adverb, multipliers for timing were applied. The adjustments to timing would compress or lengthen the amount of time between keyframes. Arm joint positions were adjusted, changing the overall motion paths for the wrists. Rotational adjustments were made to the wrists to expand expressivity.

However, adjusting joint rotations and joint positions are not sufficient to convey the adverbial changes. Ease-in and ease-out were incorporated to increase the perceived naturalism (Thomas and Johnston 1995). Ease-in refers to the Slow In principle and ease-out refers to the Slow Out principle found in animation. A fast ease-in can model the sudden breaking on a car. A slow ease-out can simulate a gentle acceleration. Adding these allows for changes to the speed of the curve of an animation (Burtnyk & Wein, 1976). Ease-in gives a slow start to the transition and ease-out gives a slow end.

The model was extended to include the proximal joints of the spine to alleviate the need of requiring an unnatural exaggerated extension of the arms as the path of the wrist moved further from the. body. This aligns with the migration of motion between distal and proximal joints as described by Brentari (1998).

6. Validating the Model

To evaluate for generalizability, the model has been further tested on several other verbs. Testing the limitations of the model required a selection of a variety of verb signs. Selected signs included use of one or two hands, varying use of contact or noncontact, and repeating and nonrepeating motion. Selection is shown below in Table 1. Table 1 includes the sign used for model development, WALK, for comparison.

Sign	*Two- Handed*	*Contact*	*Repeating*
WALK	X		X
ASK			
GIVE		X	
THINK			X
BREATHE	X	X	X
INFORM	X		
CLEAN	X	X	
SEARCH	X		X

Table 1: Shows chosen signs and their characteristics for selection.

It was also important to evaluate signs that displayed a different motion path than that used for the model. A neutral version of WALK alternates the extension and retraction of the arm, led by the hand, in a flat motion path that does not vary up or down. The signs chosen for further consideration of the model were again chosen for their variety of motion paths in comparison to WALK.

7. Adverbs of Manner - Speed

The adverbial modification for *slowly* proved to be the easiest to transfer to the new signs, followed by *quickly*. The changes to this set of contrasting adverbs of manner relied less on a change in motion path to convey the modification and even less on nonmanual signals, though wrist rotations were a contributing part of the model for both. Whereas the model for *slowly* was perceived across all evaluated signs, *quickly* was perceived on signs that had repeating motion, such as BREATHE, and was harder to perceive on non-repeating signs such as ASK and GIVE. This is primarily due to the short duration for the neutral version of the nonrepeating signs. When the model for *quickly* was applied, the overall timing of the sign was not changed significantly and further increasing the speed only made the animation come across as less natural. To account for this, the model was further extended to the surrounding signs. In the case that the sign where the adverbial modifier for *quickly* was applied, the prior and following signs were also modified. Figure 4 below shows an overlay of the neutral animation for BREATHE and the applied *slowly* model. Figure 5 below shows the comparison with *quickly*.

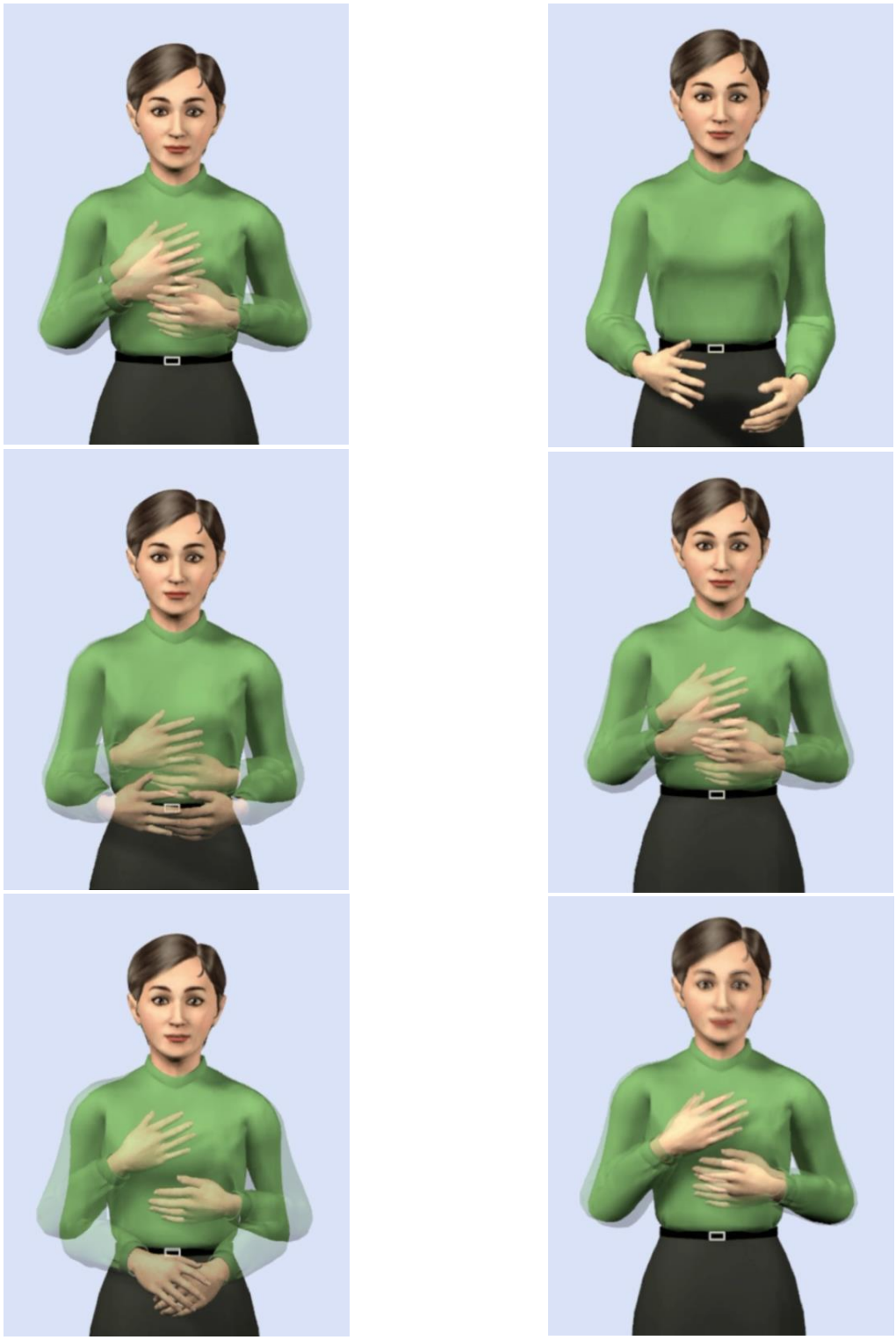


Figure 4: Shows an overlay with the neutral animation of BREATHE and BREATHE slowly.

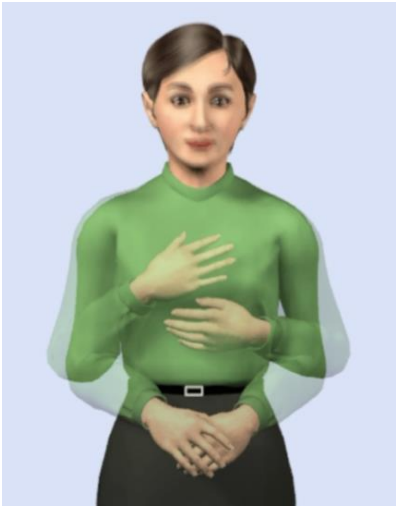


Figure 5: Shows an overlay with the neutral animation of BREATHE and BREATHE quickly.

8. Adverbs of Manner – Affect

The changes to this set of contrasting adverbs of manner relied heavily on a change in motion path to convey the modification and even more on nonmanual signals. For both *happily* and *sadly*, the timings were extended and ease-in and ease-out was used. The significant difference between the two came from the changes to the motion path. For *happily*, the model lifted and expanded the motion path. Figure 6 shows an overlay of the neutral GIVE and GIVE with the *happily* modification. For *sadly*, the model lowered and compressed the motion path. Based on the initial data collection, *sadly* showed to have a continuous lowering effect on the signing space in signs with repeating motion. This is demonstrated in Figure 7. To incorporate this into the model, keyframe data was compared in the neutral animation to see if the sign comes back near the starting position, any keys after would have an increased drop in their position and an even slower timing adjustment..

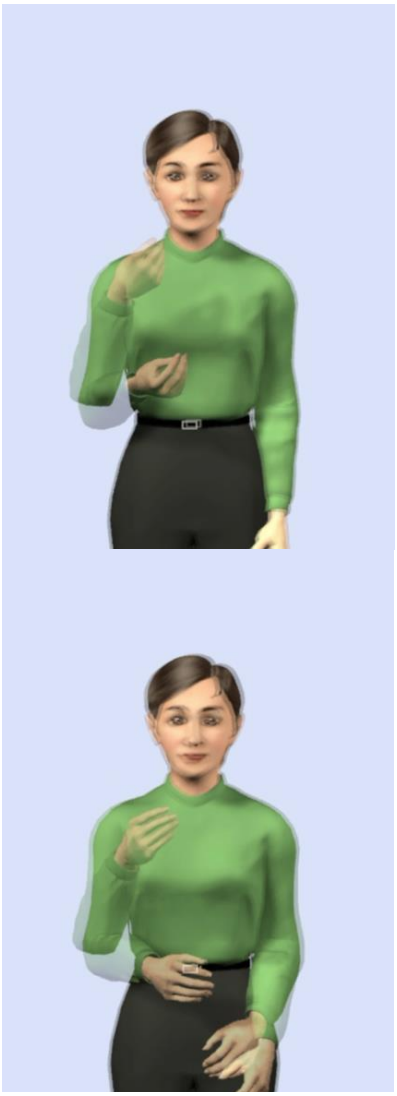


Figure 6: Shows an overlay with the neutral animation of GIVE and GIVE happily.

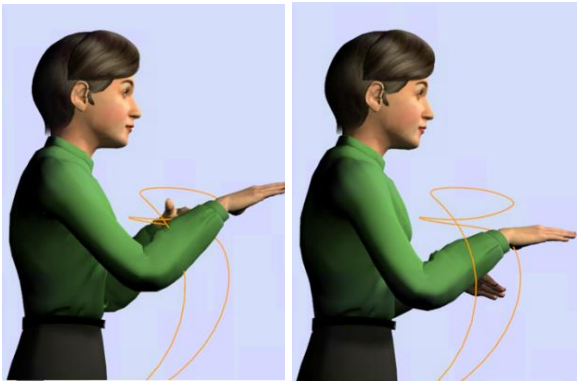


Figure 7: Shows the motion path the right wrist for WALK *slowly*. The first extension is shown higher, with the repition of the extension shown lower.

9.

9. One Handed vs Two Handed Signs

The model needed to recognize if a sign was one or two handed. Since the original model was built based on a two-handed sign, initially both arms were being modified when applied to one-handed signs. This would cause the left hand to move when there should not be movement. The model now recognizes if the sign is one-handed or two-handed based on comparison between keyframes throughout the sign. It then recognizes which hand to apply the modification to.

10. Single vs Repeating Motion Signs

When applying the model to single motion signs, there is less overall movement to aid in the perception of the adverbial modification. To help with the perception of the adverbial modifiers of intensity, surrounding signs can be adjusted. In the case of affect, the use of nonmanual signals (facial expressions) will play an important role.

11. Contact Signs

When confronted with signs that had some form of contact, GIVE, BREATHE, and CLEAN, the model did not initially take this contact into account. In several cases this resulted in the hands overshooting and ending in a collision with the body at the point of contact. To account for this, the neutral animations needed to be adjusted to include a tag on the keyframe with the contact. With this tag in place, the model would negate any positional/rotational motion path modifications applied to that keyframe. This would allow for the contact to occur as originally animated, with the surrounding keyframes being adjusted.

12. Conclusion and Future Work

To generalize the model for adverbial modifications, it did require extension for application to various signs to adjust for contact, expanding the model to the prior and trailing signs based on the adverbial modification for *quickly*, and further lowering of motion path for repeating motion based on the adverbial modification for *sadly*.

The next step in this work is to conduct a user study to test whether using multiple channels will increase the intensity of the perceived adverb. In other words, is the adverb *happily* portrayed by motion modification and nonmanual signal perceived as *more happy* than when the adverb is portrayed by motion modification alone.

13. Bibliographical References

Baker, C. & Cokely, D. (1980) American Sign Language: A teacher's resource text on Grammar and Culture. Silver Spring Md.: T.J. Publishers. Reprint, Washington, D.C.: Gallaudet University Press

Barraquand, J. & Latombe, J. (1991) Robot motion planning: A distributed representation approach. The International Journal of Robotics Research. 10.6. pp. 628-649.

Brentari, D. (1998) A prosodic model of sign language phonology. MIT Press.

Burtnyk, N. & Wein, M. (1976). Interactive Skeleton Techniques for Enhancing Motion Dynamics in Key Frame Animation. Communications of the Association for Computing Machinery, Vol 19. No. 10. pp. 564-569.

Chi, D., Costa, M., Zhao, Z. & Badler, N. (2000) The EMOTE model for effort and shape. Proceedings of the 27th annual conference on Computer graphics and interactive techniques. ACM Press/Addison-Wesley Publishing. pp. 173-182.

Eisenbeis, R.A. & Avery, R.B. (1972) Discriminant analysis and classification procedures: theory and applications. Lexington, MA: DC Heath.

Hartmann, B., Mancini, M. and Pelachaud. C. (2005) Implementing expressive gesture synthesis for embodied conversational agents. International Gesture Workshop. Berlin: Springer. pp. 188-199

Kleinsmith, A. & Bianchi-Berthouze, N. (2012) Affective body expression perception and recognition: A Survey. IEEE Transactions on Affective Computing 4, 1: pp.15-33.

Malala, V.D., Prigent, E., Braffort, A. & Berret, B. (2018) Which Picture? A Methoology for the Evaluation of Sign Language Animation Understandability.

McDonald, J., Wolfe, R., Alkoby, K., Carter, R., Davidson, M. J., Furst, J., Hinkle, D., Knoll, B., Lancaster, G., Smallwood, L., & Ougouag, N. (2005) Achieving consistency in an FK/IK interface for a seven degree of freedom kinematic chain. Proceedings of the 13th international conference in central Europe on computer graphics, vizualization and interactive digital media. pp. 171-179.

McDonald, J., Wolfe, R., Wilbur, R. B., Moncrief, R., Malaia, E., Fujimoto, S., Baowidan, S., & Stec, J. (2016). A new tool to facilitate prosodic analysis of motion capture data and a data-driven technique for the improvement of avatar motion. In Proceedings of Language Resources and Evaluation Conference (LREC). pp. 153-159.

Moncrief, R. (2019) A Model for Animating Adverbs of Manner in American Sign Language. Talk presented at the 6th Workshop on Sign Language Translation and Avatar Technology. Hamburg, Germany

Padden, C. (2016) Interaction of morphology and syntax in American Sign Language. Routledge.

Schnepp, J.C., Wolfe, R.J., McDonald, J.C., & Toro, J.A. (2012) Combining emotion and facial nonmanual signals in synthesized American Sign Language. Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility.

Thomas, F. & Johnston, O. (1995). The illusion of life: Disney animation. Disney Editions.

Valli, C. & Lucas, C. (2000) Linguistics of American sign language: an introduction. Gallaudet University Press, 2000.

Zhao, L., Kipper, K., Schuler, W., Vogler, C., Badler, N. & Palmer, M. (2000) A machine translation system from English to American Sign Language. In *Conference of the* Association for Machine Translation in the Americas, pp. 54-67. Springer, Berlin, Heidelberg

Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 157–164
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

From Dictionary to Corpus and Back Again – Linking Heterogeneous Language Resources for DGS

Anke Müller, Thomas Hanke, Reiner Konrad, Gabriele Langer, Sabrina Wähl
Institute of German Sign Language and Communication of the Deaf
University of Hamburg, Germany
{anke.mueller, thomas.hanke, reiner.konrad, gabriele.langer, sabrina.waehl}@uni-hamburg.de

Abstract

The *Public DGS Corpus* is published in two different formats, that is subtitled videos for lay persons and lemmatized and annotated transcripts and videos for experts. In addition, a draft version with the first set of preliminary entries of the DGS dictionary (*DW-DGS*) to be completed in 2023 is now online. The *Public DGS Corpus* and the *DW-DGS* are conceived of as stand-alone products, but are nevertheless closely interconnected to offer additional and complementary informative functions. In this paper we focus on linking the published products in order to provide users access to corpus and corpus-based dictionary in various, interrelated ways. We discuss which links are thought to be useful and what challenges the linking of the products poses. In addition we address the inclusion of links to other, older lexical resources (LSP dictionaries).

Keywords: dictionary, corpus, cross-linking

1. Introduction

The DGS-Korpus project is a long-term project (2009-2023) that has three major aims: a) compiling a reference corpus of German Sign Language (DGS), b) publishing part of the annotated corpus, c) compiling and publishing a corpus-based dictionary DGS–German. Data collection took place from 2010 to 2012 and captured near-natural DGS data from 330 informants coming from all over Germany (Nishio et al., 2010). The *DGS Corpus* contains about 560 hours of DGS signing. Lemmatising and annotating is done with `iLex`[1] (Hanke, 2002; Hanke and Storz, 2008), a lexical database and annotation tool designed for a multi-user environment. A subset of about 50 hours was selected for publication. This *Public DGS Corpus* was published on two different portals, *MY DGS*[2] and *MY DGS – annotated*[3]. The corpus-based dictionary *Digitales Wörterbuch der Deutschen Gebärdensprache* (*DW-DGS*) is still in the making. Its final version is to be published end of 2023. In order to test and discuss form, content, and usability with the language and the research community, we make a pre-release of dictionary entries available[4]. Since the *DW-DGS* and the *Public DGS Corpus* are closely related, it is obvious to make the relation tangible for the users of both *DW-DGS* and *Public DGS Corpus*. In addition, we want to integrate information on DGS signs that was published earlier in several LSP (language for specific purposes) dictionaries German–DGS. Thus, several features link dictionary, corpus, and heterogeneous DGS language resources.

2. Data Structure and Language Resources

2.1. Data Structure

In iLex, types are database entities with unique IDs, which tokens are linked to. A type is an abstract unit of the language with a specific form that – for iconically motivated signs – is associated with a specific underlying image (König et al., 2008). Its form can have several realisations in actual use and it can have a number of different conventional meanings. In order to group tokens according to these conventional meanings, we implemented a type hierarchy (type levels) and double glossing: Each type (parent) can have one or several subtypes (children) (Konrad et al., 2018; Langer et al., 2016). At the beginning of the lemmatisation of the *DGS Corpus* data two additional type levels – qualified types and qualified subtypes (Konrad et al., 2012) – were implemented to group recurrent form variations and modifications of types or subtypes. Tokens are matched either to a type, a subtype or a qualified type. A type entity in iLex is defined at least by a gloss and a citation form in HamNoSys[5] (Hanke, 2004). Type and subtype glosses are given in *MY DGS – annotated*, whereas qualified types are used but internally in the *DGS Corpus*.

When the DGS-Korpus project started, iLex already comprised a large number of type entities and lemmatised tokens of collected data as well as of studio reproductions of isolated signs (citation form). For the production of LSP dictionaries (see Section 2.4) quite an amount of supplementary production data were available.

As before, the *Public DGS Corpus* (Section 2.2) is produced from the data stored, managed and prepared in iLex. This also applies for the *DW-DGS*. The data includes types selected for dictionary entries, studio reproductions for representing the signs' citation forms, and video sequences taken from the *DGS Corpus* to serve as examples for sign senses described in the respective entry (Langer et al., 2018).

One of the first steps when compiling a dictionary is to define which data from the corpus is to be covered by and described in a dictionary entry, that is, which types or parts of a type structure should be included. This step is called

[1] `https://www.sign-lang.uni-hamburg.de/ilex/`
[2] `http://meine-dgs.de`
[3] `http://ling.meine-dgs.de`
[4] `http://dw-dgs.meine-dgs.de`
[5] `http://www.dgs-korpus.de/index.php/hamnosys-97.html`

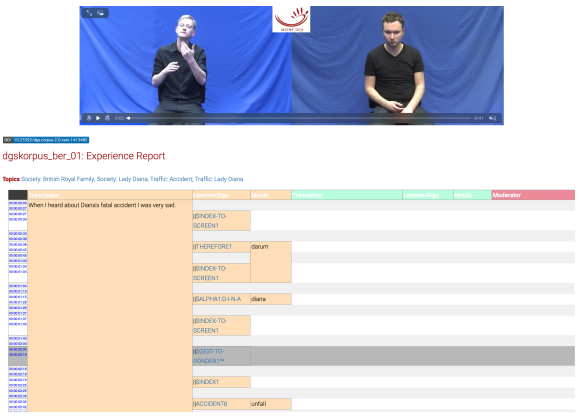


Figure 1: Online transcript view. Location: Berlin. Format: Experience Report. Topics: Society – Lady Diana.

lemma establishment (Svensén, 2009, p. 94). Rules for lemma establishment and annotation guidelines (for token-type matching, i. e. lemmatisation) serve different purposes, may follow different rules and reflect different stages of analysis (Langer et al., 2016). Lemma establishment decisions reflected in the scope of dictionary entries may not necessarily lead to changes in the data structure of type entities in iLex. Thus, a dictionary entry can cover more than one type or a type in iLex can be split up into more than one dictionary entry. It is also possible that a branch of a type structure in iLex is described in a separate entry together with data from another type. While this is unproblematic for the dictionary as a stand-alone product, it makes interlinking of corpus and dictionary more challenging.

2.2. The Public DGS Corpus

The prerequisite for building the *DGS Corpus* was the consent of the informants to collect, analyse and publish their data. In order to give something back to the language community and to make data accessible for the sign language research community, one project goal is the publication of about 50 hours of signing with annotations in German and English. We decided to publish this *Public DGS Corpus* in two portals suited for the different needs of the language and the research community. Selection, processing steps, data formats and features of Release 1 are reported in Jahn et al. (2018), for the changes in Releases 2 and 3 see Hanke et al. (2020). With Release 2, the target quantity of 50 hours was reached.

2.2.1. Portal *MY DGS*

This portal addresses users who are interested in the content of discussions, conversations, and narratives on history, life and culture of Deaf people. It contains over 47 hours of videos with translations from DGS to German as optional subtitles. *MY DGS* provides a low-threshold access to the data. The metalanguage used for description on this website is German only. In addition, 2.4 hours of jokes as part of the (German) deaf culture can be browsed (without subtitles). The videos can be filtered for 13 regions, 4 age groups, 8 formats (elicitation tasks) and 38 main topics. For the following, the 47 hours of discussions and conversations are of interest as there are links from the *DW-DGS* examples to these sequences (see Section 3.2).

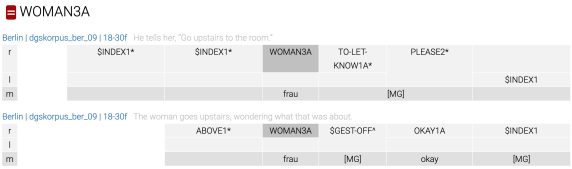


Figure 2: KWIC concordance with tokens of the subtype WOMAN3A.

2.2.2. Portal *MY DGS – annotated*

This portal made for the research community includes the video material of *MY DGS* lemmatised and annotated (except jokes) and is fully available in both English (except mouthings) and German. Of the tasks not included in *MY DGS* additional 1.7 hours of video were selected to exemplify the whole range of tasks covered in the data collection. We considered this material more important for research than for the general public. Only the tasks "Sign Names" and "Isolated Items" are not part of the *Public DGS Corpus*. In the following, we focus on the online transcript view and the types list of *MY DGS – annotated*, for detailed information on data formats and features see Hanke et al. (2020).

The online transcript shows a video with both informants in the frontal camera perspective (during the elicitation they were sitting face to face). Beneath, versioning (DOI), video name (with location/region and task) and topics are given, followed by a vertical transcript as shown in Figure 1.

All glosses in the "Lexeme/Sign" tier are clickable leading to the corresponding type or subtype section of the type entry. Figure 2 shows the section of the subtype WOMAN3A where all tokens of this subtype in the *Public DGS Corpus* are listed as a KWIC concordance (keyword in context) and highlighted by a dark grey shade: Each token is listed by its metadata and the English translation of the utterance it is part of, that is for the first token of WOMAN3A: Berlin | dgskorpus_ber_09 | 18-30f, "He tells her, 'Go upstairs to the room.'" The translation tag limits the range out of which the left and right neighbours of the target token are taken. That is why some key tokens show less than three neighbour tokens left or right. All glosses of the neighbouring tokens are clickable leading to the respective type entry. Below the token glosses annotated mouthings or mouth gestures are shown.

The parent type of WOMAN3A is EARRING1A^, which is listed at the head of this type entry (see Figure 3). In case that a studio reproduction of the citation form of the sign is available, the video is displayed under the gloss name. Studio reproductions made for the *DW-DGS* show the isolated sign in four perspectives. If the video is taken from prior productions, only one perspective is given. In the course of the production of dictionary entries more and more videos will be added.

2.3. Corpus-based DGS Dictionary

The *DW-DGS* is based on the total of annotated material of the *DGS Corpus* (with over 601700 tokens), which exceeds

Figure 3: Type entry start of EARRING1A^ with video (citation form) and links to *DW-DGS* entry 156 and same type in the gloss index of LSP dictionary *Health & Nursing*.

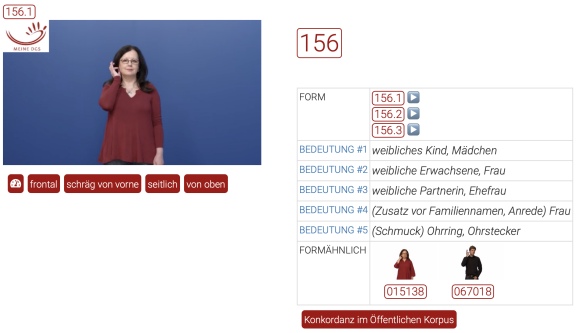


Figure 4: Entry 156 with three form variants, overview of senses with sign posts and two cross-references as micons.

the published data (more than 373800 tokens). The *DW-DGS* aims at the description and documentation of signs as they are used in everyday signing, as represented in the corpus data. Though it serves the function of a bilingual dictionary with German translational equivalents and an index of German, the focus is on the description of DGS and its structures independent of German, as if in a monolingual dictionary.

The *DW-DGS* addresses diverse user groups including the language community and native signers as well as beginning and advanced learners, the general public as well as linguists. The pre-release is an incremental publication of entries along with a growing macro-structure as for example background information and search facilities. What is of interest for this paper is the structure of entries, the DGS index and the German index. The DGS index displays all entries that are fully edited by way of a micon (moving icon). One of the main design decisions for the dictionary was not to represent signs by glosses, but to use thumbnail videos and numbers instead, resulting in micons consisting of a posed still of a signing model plus a unique identification number. This prevents the user from mistaking gloss names for meaning or to confuse glosses with German, especially as German is the metalanguage for sign descriptions within the entry. The dismissal of glosses for the *DW-DGS* entries has the further advantage of avoiding a clash or discrepancy of glosses between dictionary and corpus which would occur whenever the lemma establishment does not match the lemmatisation of types in iLex. Figure 4 shows a sign entry as it appears when accessed via the DGS index. A sign entry is identified by the identification number and the citation form of the sign. Information given on a sign includes form variants of the sign, information on regional distribution, cross-references to signs with identical citation form (homonyms) and signs with similar citation forms. The main body consists of the description of the sign's senses based on the analysis of corpus data. Figure 4 shows the overview of 5 senses indicated by sign posts; each, when clicked, reveals a table of detailed information on a sense such as an explanation of meaning or usage, typically co-occurring mouthings, German translational equivalents, authentic examples directly taken from the corpus for attesting and illustrating senses, cross-referenced synonyms and antonyms, and collocational patterns.

All information given in DGS can be viewed in the fixed display window, that is, the form variants of the lemma, all signs represented as micons, and examples. Micons are used for cross-references within the dictionary – when clicking the still, the corresponding film can be viewed in the film display window; the number serves as a link to the corresponding entry. A preliminary design feature is the automatic generation of entries, if there is a cross-reference to an entry that does not exist as a fully edited article. Such an automatically generated entry shows the sign form and a link back to all entries referring to the sign in question. These back links are labeled according to their relation kind, e. g. synonym of X.

The German index is a list of translational equivalents followed by entry identification numbers giving access directly to the corresponding senses indicated by the number of the sense within an entry, e. g. entry 59#2.

In the process of manually performed sense discrimination, not every token of a type is viewed and analysed, but only a critical mass to attest or confirm the most typical senses. Particularly if a sign type has many tokens, they cannot all be reviewed in detail. Moreover, not all tokens can be assigned to the senses identified, depending on the granularity of the senses. This is why, in the DGS-Korpus project, we do not have a full sense-tagging. There is no automatic solution for a reliable sense-tagging at sight. This fact has implications on the linking of dictionary and corpus (see Section 3.2).

2.4. LSP Dictionaries German–DGS

Lexicographic work on DGS was conducted at our institute previous to the DGS-Korpus project. Between 1993 and 2010, six LSP dictionaries (*Psychology*, *Joinery*, *Home Economics*, *Social Work & Social Pedagogics*, *Health & Nursing*, and *Horticulture & Landscaping*) were compiled (Konrad, 2011; Konrad and Langer, 2012). Within the context of these projects experience, methodology, know-how, and technical tools were developed and improved.

Except for the first project, DGS equivalents in the elicited answers to words and/or picture prompts and semi-structured interviews were lemmatised and annotated using annotational tools developed at our institute. The LSP dic-

tionaries are bidirectional in that they consist of two kinds of entries – concept entries with definitions headed by the German technical term, and additional sign entries of simplex signs used in the DGS equivalents of German technical terms. These signs were listed and described in sign entries accessible through sign indexes or from cross-references within the concept entries. All entries and indexes were produced directly from the information stored, corrected and prepared in a lexical database (GlossLexer Hanke et al. (2001), then iLex). In order to make the respective sign index consistent and the numbering gapless, production glosses with continuous numbering within each product partly replaced the iLex-internal glosses. As a result, glosses for the same sign may differ between the LSP dictionaries and iLex.

When the DGS-Korpus project started, iLex already comprised a large number of type entries, lemmatised tokens, annotated mouthings/mouth gestures from data collected in previous dictionary projects as well as production data and lemmatised studio reproductions of citation forms. While information on types and therefore their description in the database may have changed over time through new data, re-evaluation of data, change of annotation conventions, or corrections, there is still a considerable number of types that are used in the *DGS Corpus* data as well as in the data of previous projects. This common base of type entries can be utilised to link from entries in the types list of *MY DGS – annotated* as well as from *DW-DGS* entries to the corresponding types in the sign entries of three LSP dictionaries: *Social Work & Social Pedagogics* (Hanke et al., 2003), *Health & Nursing* (Konrad et al., 2007), and *Horticulture & Landscaping* (Konrad et al., 2010).

3. Linking Corpus and Dictionary

3.1. Challenges

Linking *MY DGS – annotated* and *DW-DGS* entails challenges that need to be considered. First, the user groups are rather diverse with different needs. The dictionary aims at a broad public interested in DGS including researchers, whereas the research portal is aimed at a scientific public. Second, as the research portal provides transcripts it also displays glosses used for lemmatisation. Within the dictionary glosses are not used to refer to signs, micons combined with numbers are used instead. These different styles may be confusing for users. Third, as Langer et al. (2016) pointed out, lemmatisation decisions in the database do not necessarily match lemma establishment in the dictionary. Hence different types from the database appearing in the *Public DGS Corpus* types list may be mapped onto one entry, or one type may be mapped onto several entries.

3.2. From Dictionary Entry to Corpus

Compiled entries of the dictionary are based on corpus occurrences. While a dictionary entry sums up forms, properties, meanings and uses of a sign, a corpus presents the data in a structured way, e.g. through a listing of all occurrences of a type and links to the source texts in annotated transcripts. The DGS-Korpus project makes both available – the results of lexicographic analysis and a structured view of tokens of the same type, which is presented as a KWIC concordance. This presentation allows users to have a look at the context a sign occurs in, as well as a comparison of left and right neighbours (for a detailed description of the KWIC concordance see Hanke et al. (2020)).

Entries in the pre-release of the *DW-DGS* contain a red button at the bottom (cf. Figure 4 or the box 'DW-DGS' in Figure 5), which when clicked opens a KWIC concordance of the tokens of all types and subtypes that constitute the respective entry, given that they occur in *MY DGS – annotated*. The view of this entry generated concordance differs from the view when accessed within *MY DGS – annotated* in some points: The list is headed by the identification number of the entry the KWIC concordance belongs to, which serves as a direct back link, and there are neither a studio reproduction nor type and subtype glosses as headings that indicate the gloss hierarchy of the iLex database (cf. box 'KWIC1' as opposed to the boxes 'KWIC2', 'KWIC3', 'KWIC4' in Figure 5). Otherwise, the same information and link structure is given with respect to the single type occurrences (tokens), that is, there is a link heading each KWIC line to the token in the respective transcript, and neighbouring glosses of the target gloss link to their respective type in another KWIC concordance (cf. arrows from KWIC1 and KWIC3). But, and this is necessarily so, the target gloss also links to the respective type in a KWIC concordance of the *MY DGS – annotated* style (KWIC3). This way a user can find out which type a particular subtype gloss may belong to.

The KWIC concordance as generated from a dictionary entry reflects the lexicographic lemma establishment, which sometimes results in sampled concordances made up from two or more types, or may also cut off a sub-branch of a type. Ideally spoken, a linguistic expert could make up their own dictionary entry by viewing all listed tokens.

Coming from the dictionary where signs are represented as stills, micons or video, the user is confronted with the use of glosses in the KWIC concordance, which they cannot directly associate to the lemma sign of the entry they may come from. If they click onto different key tokens marked by dark grey background, eventually they open all type concordances from the corpus and recognise the shown variants in the studio reproduction on top of each list, as well as the entry number of the *DW-DGS* appearing there. Though at first potentially confusing, the availability of a sampled KWIC concordance offers a lot of additional examples with a broad range of information on sign forms (modifications and phonetic variants), use and senses in different contexts, which may also include uses that are not described in the entry because they are used in a productive and sense-expanding way, or because there is too little evidence for a conventionalised use. Even the examples used in the entry may be discovered; a marking of those is a planned feature for future releases. Here, users may observe differences of segmenting and translation, which is due to our preparing an example to serve as a good example of a sense even out of context, which sometimes requires to adjust the translation of an utterance (cf. Langer et al., 2018). These adjustments are always true to the original. The examples of sign uses displayed in the KWIC concordance are not grouped according to the senses defined and listed in the

corresponding entry because tokens are not systematically sense-tagged in the corpus.
As stated above, in the pre-release of the *DW-DGS* there are many automatically generated entries without proper lemma establishment or form and sense descriptions. But they all offer the link to the corpus KWIC concordance, so a user of the dictionary can gather more information on a sign they were referred to by a cross-reference, be it a type or a subtype. Another kind of external link implemented in the dictionary entry structure is from an authentic example shown as a cut-out within the entry to the source text of the very example. Whenever an example is taken from the *Public DGS Corpus*, two red buttons show up below the video display window (see Figure 5). The first button takes the user to the beginning of the source text in *MY DGS*, where they can view the whole discourse context in full detail and observe the use of the sign of that sense in this specific case. The second button targets the beginning of the example utterance in the respective transcript of *MY DGS – annotated*.

3.3. From Corpus to Dictionary Entry

The main route leading from the *Public DGS Corpus* to the *DW-DGS* is the KWIC concordance showing all the occurrences for one type and the dependent subtypes. If there is a studio reproduction of the sign's form available, it is displayed under the gloss of the type. Next to that video you may find one or more entry numbers linking to the dictionary, if there is an entry already in existence. The number of entries linked to a type depends on lemma establishment decisions (Section 2.1) that do not necessarily map 1:1 to the type structure. Thus there are three different cases of mapping between corpus and dictionary. The simplest case is a 1:1 mapping between sign type and dictionary entry. If an entry comprises several sign types, e. g. because they are phonological variants of one another, the mapping is 1+n:1 from corpus to dictionary (see box 'KWIC2' and 'KWIC3' in Figure 5). The third case is that a subtype is defined as an entry in its own right compared to the rest of the type, e. g. because it is a sign modification with a specific meaning the other forms of the sign do not show. In that case the mapping is 1:1+n (see box 'KWIC2'). Naturally, confusion may occur especially with the third case, so information on the project's lemma establishment principles are needed in order to make the decisions transparent. The benefit for the users is that they may find information on a sign's possible meanings and uses that are not provided via the types list and concordance view directly. The dictionary also features prepared information on e. g. collocations of the sign.

4. Linking to Heterogeneous Resources

4.1. Challenges

The *Public DGS Corpus* and the *DW-DGS* are complementary products that are both based on the same data collected and are created in parallel with relation to each other and in the same time span with interlinking planned from the very beginning. A different case is the linking to previously published lexical resources, namely the LSP dictionaries *Social Work & Social Pedagogics*, *Health & Nursing*, and *Horticulture & Landscaping*.

When comparing these to the *DGS Corpus* and *DW-DGS*, several important differences can be observed:

- They cover specialised language and were aimed at sign expressions of technical terms as opposed to everyday language in *DGS Corpus* and *DW-DGS*.
- The main portion of the data collection involved elicitation of isolated signs for technical terms following a German word list as opposed to natural signing in context. Answers consist of a demonstration of the respective signs and do not include their actual use in a linguistic context, a prerequisite of analysing usage.
- Due to the elicitation method it was not always completely clear which of the answers were established signs and which were spontaneously made up translations such as loan translations, homophone calques and productive signs (cf. König et al., 2008, p. 380). For an evaluation and selection of the signs to be shown in the dictionaries, native speakers' intuition of Deaf team members and the recurrent use by several informants were used as criteria.
- Methodological and technical aspects of elicitation, annotation and production were according to the standards of the respective time. This means that the quality of contents and lemmatisation may be somewhat outdated in comparison to today's standards and rules.

Although the data of the LSP dictionaries are stored and maintained in iLex, it happened for several reasons that IDs used for type entries in the gloss index of an LSP dictionary changed or got lost. In these cases the IDs have to be reconstructed or a mapping with actual type IDs needs to be done manually.
For the joint German index the challenge was to come up with a feasible rule to filter out links to LSP sign entries that were already covered by *DW-DGS* entries.

4.2. Rationale for Linking to Older Resources

The *Public DGS Corpus* and *DW-DGS* are intended to become the preferred reference tools for information on DGS when finished. Since they are online products they can be interconnected with each other and with other lexical resources of DGS and can thus serve as a common gateway also to these other resources. Resources can be linked without too much extra cost when the technical matching of sign entries to the entries of the respective resources can easily be achieved, when there is no legal problem with access rights and it can be ensured that the other resources will be unchanged and stay available in the future (sustainability). All these conditions are fulfilled for the LSP dictionaries in question. Reasons for linking are:

- Linking from the *MY DGS – annotated* type entries to the LSP dictionaries can easily be achieved because of shared iLex type IDs.
- Sign entries of the LSP dictionaries contain descriptions and general information on the simplex signs that were used in translations for technical terms. These

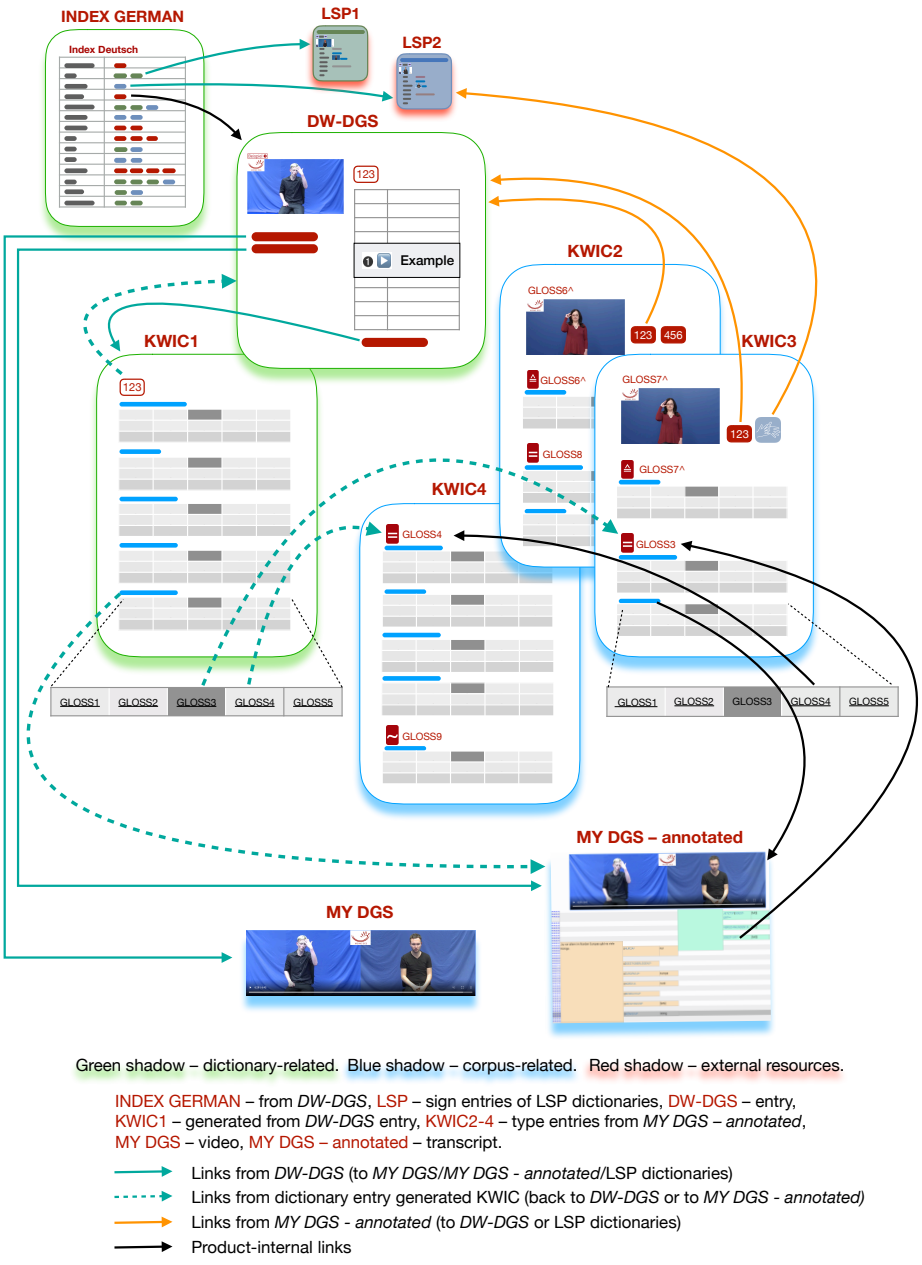


Figure 5: Implemented linking from corpus, dictionary, and other DGS resources.

signs were "[. . .] described almost as they would be in a general sign language dictionary" (König et al., 2008, p. 387). Entries include a representative movie of the citation form, identified conventional meanings and for iconic signs a description of the underlying image. This information serves the same information needs of the user as the *DW-DGS*, that is, information on the typical, everyday use of a specific sign.

- While the first entries of the pre-release *DW-DGS* are published online this resource should contain material on as many signs as possible so that a user can find at least some information when searching for a sign – even if there is not yet a fully finished corpus-based entry available. Including older information on signs that is already available and easily integrated into the resource increases the chances that a user finds useful information even at this early stage of production.

- LSP sign entries include a description of the iconic base of the signs, a piece of information not included in the *DW-DGS* entries. Making this information available can be considered as an additional gain. This is one of the reasons to link from the *MY DGS – annotated* type entries to LSP sign entries also in cases when a *DW-DGS* entry already exists.

There are two places where linking from DGS-Korpus products to the LSP dictionaries is implemented.

4.3. Linking from Corpus

MY DGS – annotated type entries link to LSP sign entries whenever a matching is available to one of the LSP products. The links are shown even if there is also a preferred link to an already existing *DW-DGS* entry. Links are done via a button representing the LSP dictionary and jump directly to the corresponding LSP sign entry (see box 'KWIC3' in Figure 5).

4.4. Linking from German Index of *DW-DGS*

The German index of the *DW-DGS* is compiled from translational equivalents provided in the entries for different senses of the described signs. German words with disambiguating context link directly to the corresponding sense in the respective entry. Not all equivalents given in the entries appear in the index. More systemic equivalents are included while less systemic equivalents (Hausmann and Werner, 1991; Héja, 2017) are excluded to avoid confusion. For those that are to appear in the index disambiguating information is added whenever the need arises to differentiate between separate senses of the German word or to distinguish between different sign senses to which the equivalents are addressed. LSP dictionary sign entries include one or several conventional meanings of the sign, realised as a German word translation sometimes with a disambiguating context added. These equivalents and contexts can be used to produce a joint German index of *DW-DGS* and LSP sign entries. *DW-DGS* translational equivalents and their disambiguating contexts are controlled for consistency while LSP translational equivalents and contexts come as they are in the product. In order to lead users to the preferred source of information – that is the corpus-based *DW-DGS* – and to avoid the confusion of multiple entries covering roughly the same scope only links to LSP sign entries are given when there is not yet a *DW-DGS* entry available.
When there is no disambiguation context given for the LSP equivalent but already existing, disambiguated *DW-DGS* equivalents, the links to the LSP sign entries are filtered out to avoid confusion and because the expectation is that *DW-DGS* sense covering might just be more detailed. However, this automatic filtering as a consequence might also filter out links to additional signs covering the same concepts or additional senses of the German word not contained in the *DGS Corpus* material and therefore not covered by the *DW-DGS* entry. In order to avoid taking out links to material not covered by the *DW-DGS* entries a manual inspection of possibly conflicting cases would be necessary to decide each case individually.

The resulting joint German index includes German words with or without a disambiguating context and links to either the *DW-DGS* entries or to sign entries of one or several LSP dictionaries (see box 'INDEX GERMAN' in Figure 5). Links to a *DW-DGS* entry appear as a red button with entry number and sense number, links to LSP entries are shown as IDs.

5. Conclusion

The DGS-Korpus project meets the vision of Kristoffersen and Troelsgård (2012, p. 99) of integrating sign language corpora and co-built dictionaries in some points. A combined product combines benefits of both a dictionary and a corpus, in addressing different user groups in various ways, providing independent use of either resource, but also close interconnection. Thus it respectively invites the language community or linguists to benefit from either the corpus or the dictionary.
With stand-alone products, there is no need to intermediate the scope of dictionary entries and the scope of type entries. In addition, as only the annotated corpus uses glosses, there is no conflict of labels. But the point of possible confusion has shifted to the places where dictionary and corpus are interlinked (see Section 3.2). This drawback is, in our view, clearly outweighed by the advantages: The interlinking documents how *DW-DGS* and *MY DGS – annotated* are built upon the same basis in a transparent way, it supports full access to resources and offers a large pool of usage examples.
Asmussen (2013, p. 1084) sets a high standard in the kind of interrelationship of what he calls a "combined dictionary-corpus product in the strict sense": Dictionary and annotated corpus "should be separately accessible" and "they should be linguistically interlinked, i. e. syntactically, semantically, and that means not only by shallow string similarities." He suggests a sense-specific linking of corpus tokens to dictionary entries (Asmussen, 2013, p. 1086). From what has been said above, a sense-tagging of the complete annotated sign language corpus is not feasible within a reasonable time. Instead, we offer a way to access from a corpus token via the referenced type to the dictionary entries. Users are able to scan the sense overview in the entry and check against the given sense definitions. For the future prospect, we think a crowd-sourcing tool that engages users to allocate tokens to the best fitting sense of the corresponding dictionary entry would be useful. These feedback inputs could be gathered, evaluated and redelivered in order to enhance the quality of KWIC concordances.

6. Acknowledgements

This publication has been produced in the context of the joint research funding of the German Federal Government and Federal States in the Academies' Programme, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies' Programme is coordinated by the Union of the Academies of Sciences and Humanities.

7. Bibliographical References

Asmussen, J. (2013). Combined products: Dictionary and corpus. In *Dictionaries. An International Encyclopedia of Lexicography – Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*, Handbooks of Linguistics and Communication Science, pages 1081–1090. De Gruyter Mouton, Berlin, Boston.

Hanke, T. and Storz, J. (2008). iLex – A Database Tool for Integrating Sign Language Corpus Linguistics and Sign

Language Lexicography. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages at LREC*, pages 64–67, Marrakech, Morocco. European Language Resources Association.

Hanke, T., Konrad, R., and Schwarz, A. (2001). GlossLexer: A multimedia lexical database for sign language dictionary compilation. *Sign Language & Linguistics*, 4(1-2):171–189.

Hanke, T., Schulder, M., Konrad, R., and Jahn, E. (2020). Extending the Public DGS Corpus in Size and Depth. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages at LREC*, Marseille, France. European Language Resources Association.

Hanke, T. (2002). iLex – A tool for Sign Language Lexicography and Corpus Analysis. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 923–926, Las Palmas, Canary Islands, Spain. European Language Resources Association.

Hanke, T. (2004). Hamnosys – Representing Sign Language Data in Language Resources and Language Processing Contexts. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages at LREC*, pages 1–6, Lisbon, Portugal. European Language Resources Association.

Hausmann, F. J. and Werner, R. O. (1991). Spezifische Bauteile und Strukturen zweisprachiger Wörterbücher: eine Übersicht. In *Wörterbücher: Ein internationales Handbuch zur Lexikographie*, Handbücher zur Sprach- und Kommunikationswissenschaft, pages 2729–2769. De Gruyter Mouton, Berlin, Boston. Reprint 2017.

Héja, E. (2017). Revisiting Translational Equivalence: Contributions from Data-Driven Bilingual Lexicography. *International Journal of Lexicography*, 30(4):483–503.

Jahn, E., Konrad, R., Langer, G., Wagner, S., and Hanke, T. (2018). Publishing DGS Corpus Data: Different Formats for Different Needs. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages at LREC*, pages 107–114, Miyazaki, Japan. European Language Resources Association.

König, S., Konrad, R., and Langer, G. (2008). What's in a Sign? Theoretical Lessons from Practical Sign Language Lexicography. In *Signs of the Time. Selected Papers from TISLR 8*, pages 379–404, Barcelona, Spain. Signum-Verlag. The International Conference on Theoretical Issues in Sign Language Research took place at the University of Barcelona between 30 September and 2 October 2004.

Konrad, R. and Langer, G. (2012). Fachgebärdenlexikographie am Institut für Deutsche Gebärdensprache. *eDITion – Fachzeitschrift für Terminologie*, 1/2012:13–17.

Konrad, R., Hanke, T., König, S., Langer, G., Matthes, S., Nishio, R., and Regen, A. (2012). From Form to Function. A Database Approach to Handle Lexicon Building and Spotting Token Forms in Sign Languages. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages at LREC*, pages 87–94, Istanbul, Turkey. European Language Resources Association.

Konrad, R., Hanke, T., Langer, G., König, S., König, L., Nishio, R., and Regen, A. (2018). Public DGS Corpus: Annotation Conventions. Project Note AP03-2018-01, DGS-Korpus project, IDGS, Hamburg University, Hamburg, Germany.

Konrad, R. (2011). Die Erstellung von Fachgebärdenlexika am Institut für Deutsche Gebärdensprache (IDGS) der Universität Hamburg (1993-2010). Revised version of doctoral thesis.

Kristoffersen, J. H. and Troelsgård, T. (2012). Integrating corpora and dictionaries: Problems and perspectives, with particular respect to the treatment of sign language. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages at LREC*, pages 95–100, Istanbul, Turkey. European Language Resources Association.

Langer, G., Troelsgård, T., Kristoffersen, J., Konrad, R., Hanke, T., and König, S. (2016). Designing a Lexical Database for a Combined Use of Corpus Annotation and Dictionary Editing. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages at LREC*, pages 143–152, Portorož, Slovenia. European Language Resources Association.

Langer, G., Müller, A., Wähl, S., and Bleicken, J. (2018). Authentic Examples in a Corpus-Based Sign Language Dictionary – Why and How. In *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, pages 483–497, Ljubljana, Slovenia. Ljubljana University Press.

Nishio, R., Hong, S.-E., König, S., Konrad, R., Langer, G., Hanke, T., and Rathmann, C. (2010). Elicitation Methods in the DGS (German Sign Language) Corpus Project. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages at LREC*, pages 178–185, Valletta, Malta. European Language Resources Association.

Svensén, B. (2009). *A Handbook of Lexicography: The Theory and Practice of Dictionary-Making*. Cambridge University Press, Cambridge, United Kingdom.

8. Language Resource References

Hanke, T., Konrad, R., Schwarz, A., König, S., Langer, G., Pflugfelder, C., and Prillwitz, S. (2003). *Fachgebärdenlexikon Sozialarbeit/Sozialpädagogik*. Arbeitsgruppe Fachgebärdenlexika, IDGS, Hamburg University, URL `http://www.sign-lang.uni-hamburg.de/slex/`.

Konrad, R., Langer, G., König, S., Hanke, T., and Prillwitz, S. (2007). *Fachgebärdenlexikon Gesundheit und Pflege*. Arbeitsgruppe Fachgebärdenlexika, IDGS, Hamburg University, URL `http://www.sign-lang.uni-hamburg.de/glex/`.

Konrad, R., Langer, G., König, S., Hanke, T., and Rathmann, C. (2010). *Fachgebärdenlexikon Gärtnerei und Landschaftsbau*. Arbeitsgruppe Fachgebärdenlexika, IDGS, Hamburg University, URL `http://www.sign-lang.uni-hamburg.de/galex/`.

Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 165–170
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

Automatic Classification of Handshapes in Russian Sign Language

Medet Mukushev*, Alfarabi Imashev*, Vadim Kimmelman†, Anara Sandygulova*
*Department of Robotics and Mechatronics, School of Engineering and Digital Sciences, Nazarbayev University
Kabanbay Batyr Avenue, 53, Nur-Sultan, Kazakhstan
†Department of Linguistic, Literary and Aesthetic Studies, University of Bergen
Postboks 7805, 5020, Bergen, Norway
mmukushev@nu.edu.kz, alfarabi.imashev@nu.edu.kz, vadim.kimmelman@uib.no, anara.sandygulova@nu.edu.kz

Abstract

Handshapes are one of the basic parameters of signs, and any phonological or phonetic analysis of a sign language must account for handshapes. Many sign languages have been carefully analysed by sign language linguists to create handshape inventories. This has theoretical implications, but also applied use, as an inventory is necessary for generating corpora for sign languages that can be searched, filtered, sorted by different sign components (such as handshapes, orientation, location, movement, etc.). However, creating an inventory is a very time-consuming process, thus only a handful of sign languages have them. Therefore, in this work we firstly test an unsupervised approach with the aim to automatically generate a handshape inventory. The process includes hand detection, cropping, and clustering techniques, which we apply to a commonly used resource: the Spreadthesign online dictionary (www.spreadthesign.com), in particular to Russian Sign Language (RSL). We then manually verify the data to be able to apply supervised learning to classify new data.

Keywords: Sign Language Recognition, Machine Learning Methods, Information Extraction

1. Introduction

Signs in sign languages are composed of phonological components put together under certain rules (Sandler and Lillo-Martin, 2006). In the early days of sign language linguistics, three main components were identified: handshape, location on the body, and movement, while later orientation and non-manual component were added. A recent paper stresses the need to combine interdisciplinary approaches in order to build successful sign language processing systems that account for their complex linguistic nature (Bragg et al., 2019).

By deploying a number of computer vision approaches, this paper aims to automate one of the most time-consuming tasks for linguists i.e. creation of a handshape inventory. Many researchers worked on establishing phonetic handshape and phonemic handshape inventories (see e.g. (Van der Kooij, 2002; Nyst, 2007; Tsay and Myers, 2009; Kubuş, 2008; Klezovich, 2019). In all of these works, handshapes were extracted and annotated manually (Klezovich, 2019).

Klezovich (2019) proposed the first handshape inventory for Russian Sign Language (RSL) by applying semi-automatic approach of extracting hold-stills in a sign video based on images overlay approach. The reason for extracting hold-stills from the rest of the video frames is due to the fact that handshapes are the most clear and visible in hold positions, and transitional movements never contain distinct handshapes. Klezovich proposed to extract hold-stills and then manually label only these frames, which can significantly speed up the process of creating handshape inventories (Klezovich, 2019).

In this paper, we test an automatic approach to generating handshape inventory for Russian Sign Language. First, we try an unsupervised learning and demonstrate that the results are unsatisfactory, because this method cannot distinguish handshapes separately from orientation and location in their classification. Second, we manually label a training dataset according to HamNoSys handshapes (Hanke, 2004), and demonstrate the utility of the supervised learning on new data.

2. Handshape as a phonological component

Ever since the seminal book by Stokoe (1960) on American Sign Language (ASL), signs in sign languages are analyzed as consisting of several parameters, one of the major ones being handshape (Sandler and Lillo-Martin, 2006). Handshape itself is not considered an atomic parameter of a sign, usually being further subdivided into selected fingers and finger flexion (Brentari, 1998).

Much research has been devoted to theoretical approaches to handshapes (see (Sandler and Lillo-Martin, 2006) for an overview), as well as to descriptions of handshape inventories in different sign languages (see e.g. (Caselli et al., 2017; Sutton-Spence and Woll, 1999; Fenlon et al., 2015; Kubuş, 2008; Klezovich, 2019; Kubuş, 2008; Van der Kooij, 2002; Prillwitz, 2005; Tsay and Myers, 2009)).

Several issues have been identified in studying handshapes that can be currently addressed using novel methods. First, many researchers identify the existence of the so-called unmarked handshapes (Sandler and Lillo-Martin, 2006, 161-162). These handshapes are maximally distinct in terms of their overall shape, they are the easiest to articulate, the most frequently occurring in signs, the first to be acquired by children, etc. For instance, in ASL, the following handshapes are generally treated as unmarked: A (fist), 5 (all fingers outstretched), 1 (index finger straight, all the other closed), E (all fingers bent and touching).

Since unmarkedness of handshapes derives from their visual and articulatory properties, it is expected that the same handshapes should be unmarked across different sign languages. This appears to be the case, although slight variation can also be observed. For instance, in Turkish Sign Language (TID), 7 handshapes can be identified as being the most frequent, including two handshapes based on the fist with or without outstretched thumb (Kubuş, 2008).

Figure 1: 135 top activated clusters for HOG descriptors.

In addition to the observation that (approximately) the same handshapes are the most frequent, a surprising finding is that the frequency of the most frequent handshapes is extremely similar across different sign languages. For instance, in British Sign Language (BSL), 50% of signs have one of the four unmarked handshapes (Sutton-Spence and Woll, 1999); in Turkish Sign Language, if we only consider the four most frequent handshapes, this would account for 57% of the signs (Kubuş, 2008), and, in ASL, the four most frequent handshapes in the ASL-LEX dataset (Caselli et al., 2017) account for 49% of all signs.

Secondly, some researchers argue that sign languages differ in their phonemic inventories, including the inventories of handshapes. For instance, Sign Language of the Netherlands has 70 phonetic and 31 phonemic handshapes (Van der Kooij, 2002), and many other sign languages are reported to have inventories of similar sizes (Kubuş, 2008; Caselli et al., 2017; Prillwitz, 2005). At the same time, Adamorobe Sign Language has been reported to have only 29 phonetic and 7 phonemic handshapes (Nyst, 2007). On the opposite end, a recent study of Russian Sign Language (RSL) based on semi-automatic large scale analysis has claimed that RSL has 117 phonetic but only 23 phonemic handshapes (Klezovich, 2019). Note however, that it is very difficult to directly compare results from different sign languages because different methods of assessing phonemic status of handshapes are used.

So we can observe both similarities in handshapes across different sign languages, as well as considerable variation. At the same time, it is difficult to make direct comparison because different datasets and annotation and classification methods are applied in different studies.

In the current study, we propose and test a method that can be applied to classifying handshapes across many sign languages using a common data set: the Spreadthesign online dictionary (www.spreadthesign.com). As a proof of concept, we analyze data from Russian Sign Language.

3. Dataset pre-processing

3.1. Dataset

The dataset was created by downloading videos from the Spreadthesign online dictionary (www.spreadthesign.com). We have downloaded a total of 14875 RSL videos from the website. The videos contain either a single sign or a phrase consisting of several signs.

Klezovich (2019) used the Spreadthesign online dictionary too, and after removing compounds, dactyl-based and number-based signs, she ended up working with 3727 signs or 5189 hold-stills.

In our case, blur images are removed using variation of Laplacian with a threshold of 350. If the variance is lower than the threshold then image is considered blurry, otherwise image is not blurry. Normally, we select threshold by trial and error depending on a dataset, there is no universal value. This reduced the number of total images from 141135 images to 18226 cropped images of hands.

3.2. Hand extraction

Hand detection can be considered as a sub-task of object detection and segmentation in images and videos. Hands can appear in various shapes, orientations and configurations, which creates additional challenges. Object detection frameworks such as MaskRCNN (He et al., 2017) and CenterNet (Duan et al., 2019) can be applied for this task.

However, occlusions and motion blur might decrease accuracy of the trained models. For these reasons, in this work, we used a novel CNN architecture namely Hand-CNN (Narasimhaswamy et al., 2019). Its architecture is based on the MaskRCNN (He et al., 2017) with an additional attention module that includes contextual cues during the detection process. In order to avoid issues with the occlusions and motion blur, Hand-CNN's proposed attention module is intended for two types of non-local contextual pooling, feature similarity and spatial relationship between semantically related entities. The Hand-CNN model provides segmentation, bounding boxes and orientations of detected hands. We utilize the predicted bounding boxes to crop hands with two padding parameters: a 0-pixel padding and a 20-pixel padding. As a result, the first group contains cropped images of detected hands only, while the other group contains cropped images of hands and their positions relative to the body.

3.3. Image pre-processing

To images with a 0-pixel-padding on detected hands, we apply Histogram of Oriented Gradients (HOG) descriptors (Dalal and Triggs, 2005). HOG feature descriptors are commonly used in computer vision for object detection (e.g. people detection in static images). This technique is based on distribution of intensity gradients or edge directions. Firstly, an image is divided into small regions and then each region has its histogram of gradient directions calculated. Concatenations of these histograms are used as features for clustering algorithm. In this work we use "feature" module of the scikit-image library (van der Walt et al., 2014) with the following parameters: orientations = 9, pixels per cell = (10,10), cells per block = (2,2) and L1 used as a block normalization method. Prior to this pre-processing, all images are transformed to grayscale and resized to 128 by 128 pixel images.

To images with a 20-pixel-padding on detected hands we utilize AlexNet (Krizhevsky et al., 2012). It is a Convolutional Neural Network (CNN) commonly used for various image processing tasks as a baseline architecture. We use only the first five convolutional layers with 96, 256, 384, 384 and 256 filters, as we only need to extract features for clustering purposes without the need for classification of images. Prior to feature extraction all images are resized to 224 by 224 pixels. CNN features are PCA-reduced to 256 dimensions before clustering.

4. Unsupervised Methodology

4.1. Clustering

We utilize a classical clustering algorithm, namely k-means. Thus, k-means implementation by (Johnson et al., 2019) is applied to ConvNet features, while scikit-learn (Pedregosa et al., 2011) implementation is applied to HOG features. Each training is performed for 20 iterations with random initialization.

We experimentally determined the number of clusters to be specified for clustering. It seemed like handshape orientation was also accounted for by the clustering algorithm, the idea was to increase the number of clusters to force the algorithm to differentiate between orientations. By trying

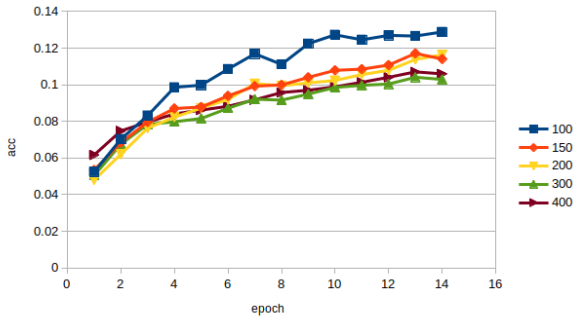


Figure 2: Average Silhouette Coefficient scores for the model trained on AlexNet features

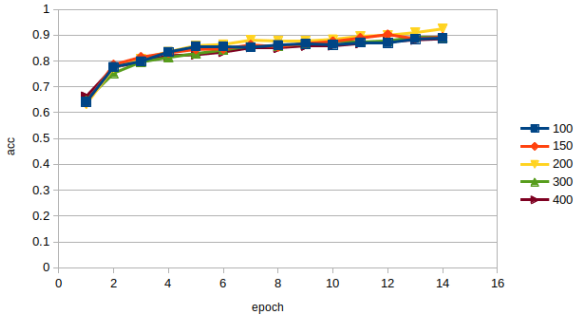


Figure 3: Normalized mutual information

varying step size, we ended up with setting for the following sizes: 100, 150, 200, 300 and 400 clusters.

4.2. Analysis and evaluation

We use two metrics to evaluate the performance of the clustering models: the Silhouette Coefficient and Normalized Mutual Information (NMI).

When the ground truth labels are not known for predicted clusters, Silhouette Coefficient score is applied. Silhouette method is used for interpretation and validation of clustering analysis (Rousseeuw, 1987). Its value gives understanding of how similar an item is to its own cluster compared to other clusters. Silhouette is bounded between -1 and +1, where a higher value means that a clustered item is well matched to its cluster and less matched to other clusters. As can be seen from Figure 2, the maximum value of Silhouette Coefficient score is observed for the model trained on AlexNet features for 100 clusters after 15 epochs. However, the score itself is just slightly over 0.12 which indicates that our clusters are overlapping.

In addition, we use predicted labels to measure NMI. It is a function that measures the agreement between predicted and actual labels. Perfect labeling gives score of +1 and bad labeling gives negative scores. As we can see from Figure 3, all models with different number of clusters result in the scores reaching 0.9 after 15 epochs.

The reason for such results might be that image descriptors for hands are too close to each other, which makes it difficult for the algorithm to differentiate. At the same time, NMI score indicates that predicted labels are almost the

same after each training epoch. In order to increase density of predicted clusters additional pre-processing of images is required.

4.3. Results

Figure 1 gives us insights about the results of applying unsupervised clustering to handshapes. First, it is clear that the algorithm does not distinguish classes only based on handshapes, but also based on orientation (for images with 0-pixel-padding), and also based on localization (for images with 20-pixel-padding). If the linguistic task of creating an inventory of phonemic handshapes is at stake, this is a clear disadvantage of this approach.

Second, despite its shortcomings, the method does provide some linguistically interesting results. Specifically, one can see that the handshapes which are expected to be unmarked (A, 5, 1) appear frequently and as labels for multiple classes. Thus, even though the classification is not linguistically relevant, the effect of markedness is still visible in the results of this unsupervised approach.

5. Supervised Methodology

5.1. Dataset

Given that the unsupervised approaches did not result in a clustering reflective of relevant handshape classes, we turned to a supervised approach. The results of HOG clustering was used as the initial dataset that contained 140 clusters of 18226 images. It was decided to manually clean the automatically generated clusters for inaccuracies. This task was performed by four undergraduate students, who divided the folders first between each other and then one person merged all of them.

First, each cluster (folder) was visually scanned for the most frequently classified handshape in order to remove handshapes that did not belong there from that folder. These steps were performed for all 140 folders. Since there were many folders of the same handshape with the only difference in orientation, they were merged, which resulted in 35 classes and a large unsorted (junk) folder. Thus, the final version of the dataset contains 35 classes of 7346 cropped images with 0-pixel-padding.

The classes were created using intuitive visual similarity as a guide, and by linguistically naive annotators. However, a post factum analysis shows that the manual classification is linguistically reasonable as an approximation of a phonological inventory. Specifically, the classes that were created are distinguished by selected fingers, spreading (spread or not), and finger position (straight, bent, curved). Thumb position is only used as a distinguishing feature for opposed thumb vs. all other possibilities. Non-selected finger position is not taken into account. This reasonably approximates features relevant for proposed phonological inventories in other sign languages, and, as such, can be used for RSL as well.

If phonetic classes were the target, then classes would also need to be distinguished by exact thumb position and also by the differences in non-selected fingers. In such a case the full inventory of possible handshapes described in HamNoSys (Hanke, 2004) could be used as the basis for manual

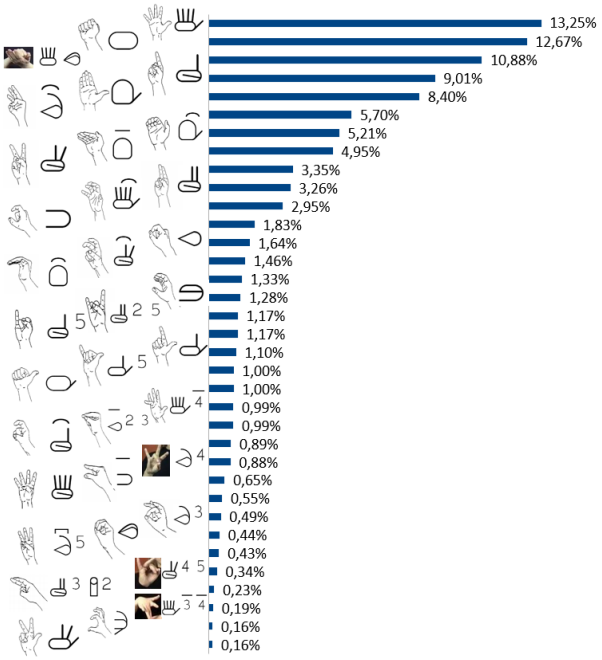


Figure 4: Handshape classes count

classification. However, the dataset we used appears to be too small to attempt a phonetic classification.

The manually labeled subset was later divided into a training set with 6430 images and a validation set with 916 images. Figure 4 shows the number of tokens for each class in a training and validation sets combined. Figure 4 also shows a linguistically relevant result: our manual classification of handshapes also demonstrates the expected frequency properties of marked and unmarked handshapes. In particular, the most frequent handshapes are the ones expected to be unmarked: A (fist), 5 (hand with all fingers spread), 1 (index finger), and B (a flat palm)).[1] These forms together constitute 48% of all handshapes (if the two-handed signs are disregarded).

5.2. ConvNet and transfer learning

Training an entire ConvNet from scratch for a specific task requires big computational resources and large datasets, which are not always available. For this reason, a more common approach is to use ConvNet that was pretrained on large datasets, such as ResNet-18 or ImageNet (which contains 1.2 million images divided into 1000 categories) as a feature extractor for a new task. There are two common transfer learning techniques based on how we use pretrained ConvNet: finetuning the ConvNet and ConvNet as a fixed feature extractor. In the first technique, we use weights of a pretrained model to initialize our network instead of random initialization. All the layers of the ConvNet are trained. In the second approach, we freeze the weights for all of the network layers and only the last final fully connected layer is changed with random weights and

[1]The *10.88%* class in Figure 4 includes all two-handed signs, which we do not attempt to classify according to handshape at the moment.

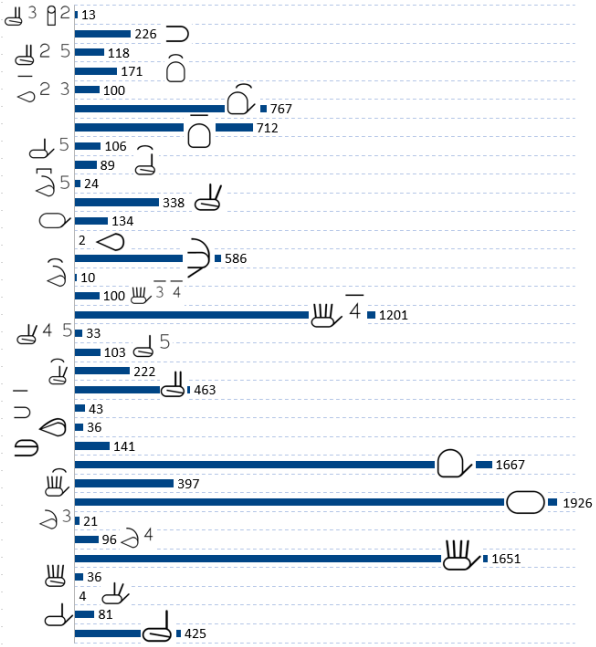


Figure 5: Handshape classes count using classifier

only this layer is trained. We implemented our networks using PyTorch (release: 1.4.0) that is an open source machine learning library (Paszke et al., 2019). Our code is based on Chilamkurthy's Transfer Learning for Computer Vision Tutorial (Chilamkurthy, 2017). ResNet-18 (He et al., 2016) model was used as a pretrained model.

5.3. Results

We trained two networks using both approaches. Each model was trained for 200 epochs. Using the second approach (i.e. ConvNet as a fixed feature extractor with only the last layer trained), the best accuracy of 43.2% was achieved. On the other hand, the first approach (i.e. finetuning the ConvNet and training all layers) demonstrated a better accuracy of 67%. Therefore, the finetuned model was used for further accuracy improvements. First, we added data augmentation to increase the number of samples. Samples were randomly rotated and visual parameters (brightness, contrast, and saturation) were randomly changed with a probability of 0.25. This helped to increase the accuracy of the best model up to 74.5% after 200 epochs. Later, we used this trained model to predict labels for all 18226 handshapes. In order to remove cases that were misclassified, a threshold for prediction probability was set to 0.7. And as a result, 12042 samples were classified. Figure 5 demonstrates the number of predicted samples for each class.

6. Discussion

6.1. Insights from unsupervised and supervised approaches

The current study shows that the unsupervised approach does not seem promising in the task of automating handshape recognition. The main problem is that the category of handshape is linguistically relevant, but not visually separable from orientation and location by this very basic data-driven approach.

We have demonstrated that an alternative approach involving a manual classification step can be quite effective. However, manual classification is problematic for obvious reasons, as it involves human judgment.

Both approaches, however, offer some linguistically relevant insights, specifically concerning unmarked handshapes. In the unsupervised approach, it is clear that many clusters are assigned unmarked handshapes as labels, which can be explained by both their frequency and visual salience. In the supervised approach, our manual classification of 7346 handshapes demonstrated that the unmarked handshapes (A, 1, 5, B) are indeed the most frequent ones. Finally, applying the ConvNet model to the whole dataset of 18226 handshapes has shown that top 3 classes are A, B, 5. Interestingly, the 1 handshape is not in the top most frequent ones. The most likely explanation is that this handshape is frequently misclassified as the handshape with middle finger bent and the other fingers outstretched (the 'jesus' handshape in the figures), which is a rare marked handshape in the manually classified dataset, but frequent in the results using the classifier.

Thus, both successful and less successful applications of machine learning methods show the importance of unmarked handshapes in RSL. It would be interesting to extend these approaches to other sign languages for comparative purposes.

6.2. Comparison with Klezovich 2019

As discussed above, Klezovich (2019) proposed the first handshape inventory for RSL by applying semi-automatic approach of extracting hold-stills in a sign video using the same dataset used here (Spreadthesign). This gives us the opportunity to compare the results of a more traditional linguistic analysis of handshape classes in RSL with the approach used in the current study.

A direct comparison is possible between Klezovich's results and the results of our unsupervised learning approaches. Both result in a classification of handshapes. However, we have demonstrated that the results of unsupervised clustering are unsatisfactory, so it cannot be used for any linguistically meaningful applications.

As for the supervised approach, both our approach and Klezovich's analysis include manual annotation, but in different ways. Klezovich manually classified handshapes into potential phonemic classes using linguistic criteria, which resulted in a large linguistically informed inventory. We manually classified handshapes based on visual similarity into a smaller number of classes, and then used this as a dataset for machine learning.

The comparison between Klezovich's and our manual classifications is not very informative, as only the former was based on linguistic criteria. Given that Klezovich's classification was not used as a training set for automatic recognition, no comparison is possible for this aspect either. This issue is left for future research.

7. Conclusion

We have shown that by deploying a number of classical machine learning algorithms, it is possible to partially automate one of the most time-consuming tasks for linguists i.e. creation of a handshape inventory, and, in addition, to investigate frequencies of various handshapes in a large data set. At the moment, it seems that unsupervised approaches cannot be used to create handshape inventories because orientation and location differences also influence clustering, and to an even greater extent than handshape itself. A supervised approach is clearly more effective, however, it requires a manual annotation component where a substantial number of handshapes is manually classified. This introduces additional problems of determining the number of classes for manual classification. Upon achieving the satisfying unsupervised clustering results, future work will focus on comparing and applying this framework to other sign languages.

8. Bibliographical References

Bragg, D., Koller, O., Bellard, M., Berke, L., Boudreault, P., Braffort, A., Caselli, N., Huenerfauth, M., Kacorri, H., Verhoef, T., Vogler, C., and Ringel Morris, M. (2019). Sign language recognition, generation, and translation: An interdisciplinary perspective. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, pages 16–31, New York, NY, USA. ACM.

Brentari, D. (1998). *A prosodic model of sign language phonology*. MIT Press.

Caselli, N. K., Sehyr, Z. S., Cohen-Goldberg, A. M., and Emmorey, K. (2017). ASL-LEX: A lexical database of American Sign Language. *Behavior Research Methods*, 49(2):784–801, April.

Chilamkurthy, S. (2017). Transfer learning for computer vision tutorial. *https://chsasank.github.io/*.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection.

Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., and Tian, Q. (2019). Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6569–6578.

Fenlon, J., Cormier, K., and Schembri, A. (2015). Building BSL SignBank: The Lemma Dilemma Revisited. *International Journal of Lexicography*, 28(2):169–206, June.

Hanke, T. (2004). Hamnosys: representing sign language data in language resources and language processing contexts. In Oliver Streiter et al., editors, *LREC 2004, Workshop proceedings: Representation and processing of sign languages.*, pages 1–6, Paris.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969.

Johnson, J., Douze, M., and Jégou, H. (2019). Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*.

Klezovich, A. (2019). *Automatic Extraction of Phonemic Inventory in Russian Sign Language*. BA thesis, HSE, Moscow.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Kubuş, O. (2008). *An Analysis of Turkish Sign Language Phonology and Morphology*. Diploma thesis, Middle East Technical University, Ankara.

Narasimhaswamy, S., Wei, Z., Wang, Y., Zhang, J., and Hoai, M. (2019). Contextual attention for hand detection in the wild. *arXiv preprint arXiv:1904.04882*.

Nyst, V. (2007). *A Descriptive Analysis of Adamorobe Sign Language (Ghana)*. LOT, Utrecht.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Prillwitz, S. (2005). Das Sprachinstrument von Gebärdenspachen und die phonologische Umsetzung für die Handformkomponente der DGS. In Helen Leuninger et al., editors, *Gebärdesprachen: Struktur, Erwerb, Verwendung*, pages 29–58. Helmut Bukse Verlag, Hamburg.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53 – 65.

Sandler, W. and Lillo-Martin, D. (2006). *Sign language and linguistic universals*. Cambridge University Press.

Stokoe, W. (1960). *Sign Language Structure: An Outline of the Visual Communication Systems of the American Deaf.* Number 8 in Studies in Linguistics: Occasional Papers. Department of Anthropology and Linguistics, University of Buffalo, Buffalo.

Sutton-Spence, R. and Woll, B. (1999). *The Linguistics of British Sign Language*. Cambridge University Press, Cambridge.

Tsay, J. and Myers, J. (2009). The morphology and phonology of Taiwan Sign Language. In James Tai et al., editors, *Taiwan Sign Language and Beyond*, pages 83–130. The Taiwan Institute for the Humanities, Chia-Yi.

Van der Kooij, E. (2002). *Phonological Categories in Sign Language of the Netherlands. The Role of Phonetic Implementation and Iconicity.* LOT, Utrecht.

van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., Yu, T., and the scikit-image contributors. (2014). scikit-image: image processing in Python. *PeerJ*, 2:e453, 6.

Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 171–176
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

Design and Evaluation for a Prototype of an Online Tool to Access Mathematics Notions in Sign Language

Camille Nadal, Christophe Collet
Université de Toulouse
Institut de Recherche en Informatique de Toulouse
Université Paul Sabatier - UT3
118 route de Narbonne
31062 Toulouse, France
Camille.Nadal@irit.fr, Collet@irit.fr

Abstract

Our project aims at giving access to pedagogical resources in Sign Language (SL). It will provide Deaf students and teachers with mathematics vocabulary in SL, this in order to contribute to the standardisation of the vocabulary used at school. The work conducted led to Sign'Maths, an online interactive tool that gives Deaf students access to mathematics definitions in SL. A group of mathematics teachers for Deafs and teachers experts in SL are collaborating to create signs to express mathematics concepts, and to produce videos of definitions, examples and illustrations for these concepts. In parallel, we are working on the conception and the design of Sign'Maths software and user interface. Our research work investigated ways to include SL in pedagogical resources in order to present information but also to navigate through the content. User tests revealed that users appreciate the use of SL in a pedagogical resource. However, they pointed out that SL content should be complemented with French to support bilingual education. The last version of our prototype takes advantage of the complementarity of SL, French and visual content to provide an interface that will suit users no matter what their education background is. Future work will investigate a tool for text and signs' search within Sign'Maths.

Keywords: Deaf Education, Sign Language, ICT in Education

1. Introduction

Sign Language (SL) started becoming a full-fledged language in few countries in the late 1980s. In France, it is only in 2005 that the Handicap Law gives a legal recognition to French Sign Language (LSF), underlining its educative, pedagogical and cultural legitimacy (Dalle, 2003). Since then, Deaf have access to education in Sign Language. However, though teaching is provided in their mother tongue (*i.e.* LSF), the majority of the pedagogical resources uses written French. Thus, an important part of the learning process relies on students' ability to read and understand French as their second language. Moreover, when learning new concepts, Deaf students have to learn the related French vocabulary in order to understand the given examples and exercises' instructions. Consequently, their knowledge of French will likely impact their learning in all disciplines. This is particularly true in science education where Deaf students have difficulties in visualising abstract concepts (Megat Mohd Zainuddin et al., 2009). Moreover, there is a real lack of vocabulary in SL for mathematics education and dissemination in higher education. In this difficult context, the online tool Sign'Maths aims at supporting Deafs in mathematics learning. The project raises the following research questions:

- How to present mathematics notions and navigate among them ?
- Does SL alone allow sufficient understanding for the user to efficiently access the definitions ? What about the user's satisfaction ?
- Is the use of SL in an educational website appreciated/preferred or disliked by Deaf users ?

We used our team experts' knowledge on education of mathematics using Sign Language and a user evaluation of the interface to address these questions.

2. Related work

In the 21st century, we have observed numerous attempts to give Deaf Community access to ICT (Information and Communication Technologies) and e-learning. Many online dictionaries for Sign Languages exist associating words of vocal language as textual and the video of the corresponding sign, like `www.sematos.eu/lsf.html`, `dico.elix-lsf.fr`, `www.spreadthesign.com`, `nzsl.nz` ... Information is organized by alphabetic order of words or by topics like *food*, *colors*, *numbers*... Search in the dictionnary can be made by word or key-word, and some propose search by phonological parameters (from (Stokoe, 1965)) : handshape and sign' location, like in (Kristoffersen and Troelsgård, 2012). In (McKee, 2017), they also propose a search by topic domains or by tags for usage status : obscene, archaic, neologism, informal and rare. These resources are meant to be used as a support to learn Sign Language. Some like in (McKee, 2017) are made for E-learning purpose. One thing worth noting about most of these dictionaries is the lack of actual definitions for the notions, except for EU funded ELIX (`dico.elix-lsf.fr`) an online dictionary for French SL learning, that provides the learner with definitions in both SL and French. These dictionnaries are not meant to learn a discipline like mathematic in Sign Language, so very few notions are presented and they don't provide definition for these notions.

In 2004, Straetz *et al.* (2004) created a bilingual web-based

learning system for Deaf adults which allowed to retrieve a German SL video translation for each text block of the page. The same year, Debvc and Peljhan (2004) provided a web-based tool that allowed the students to watch video clips of lectures along with slides and subtitles. In 2007 in Jordan, Khwaldeh *et al.* (2007) developed interactive content and interactive tools to enable "interactivity between teacher and Deaf pupils". This included online conferences and chatting rooms where pupils can communicate with teachers or with one another. Augmented Reality was first investigated by Zainuddin *et al.* as a means to support Deaf Education. They created in 2010 an AR Science book (Zainuddin et al., 2010) to help Deaf students in visualizing abstracts concepts. In 2014, Jones *et al.* (2014) studied how a head-mounted display could help pupils in learning to read by visualizing both signs and words. Another study conducted by Adamo-Villani and Anasingaraju (2017) based on the use of Augmented Reality was carried out in 2017. It focused on mathematics learning for the Deafs and used 3D holograms to translate lessons in real time.

In the light of that, a website offering education support for abstract concepts learning in SL would be an interesting approach to explore. Our project is aimed at students, teachers, and any Deaf person who wish to learn mathematics concepts. Access to the information should be easy and should not require the use of expensive or cumbersome material. As only video technology satisfies these criteria, we chose this form for educational content within Sign'Maths.

3. Presentation of the project

Sign'Maths consists of an interactive tool that gives access to mathematical notions and their definition in Sign Language. We wish to make it freely available online. It is firstly addressed to Deaf students enrolled in high school or college, or to Deafs interested in learning mathematics definition at this level of education. This tool is also addressed to teachers and interpreters who have to translate from and to SL mathematics notions taught in these years. Another set of users that must be considered is the family and relatives of Deafs. Indeed, they can be interested in learning mathematics signs to support their Deaf relative or friend in learning lessons or practising exercises. By introducing and disseminating new signs among students, teachers, interpreters and families, Sign'Maths will participate in the harmonisation of SL vocabulary for mathematics.

3.1. The project team

We have noticed that very few Deaf people come to our university whereas they are ten a year to reach the high-school degree in our town. And those that comes don't finish the first year. We can assume several reasons for this including : the loneliness of Deaf students; the lack of interpretation service (only one third of the courses are translated); or the gap between high-school and university work methods.

In order to improve this situation, we decided to start by setting up a collaboration with secondary and high school teachers and to focus on mathematics. So we assembled a team made up of a researcher in Mathematics and a researcher in Computer science (both signers) on the side of the university, a Deaf professor (PhD in mathematics) teaching in high school (Paris), two mathematics teachers (both fluent signers) in high school in classes for Deaf students (Toulouse and Lyon), a secondary school teacher (fluent signer) in a class for Deaf pupils, two Deaf teachers in primary school for Deaf pupils, and finally two Deaf students former pupils of the high-school. Teachers in primary, middle and high school have already worked together for projects on the creation of signs for educational support. This team meets every month to develop proposals for signs for mathematics notions and their definition, and to work on the software user interface. As results, they produce all the material for the website: videos of signs of mathematics notions, their definitions and sometimes examples, as well as the corresponding texts in French and some diagrams.

3.2. Prototypes and evolution of the design

The first activity carried out is the identification of users' needs. As the access to end users was very limited, the members of the team with extended experience in Deaf education identified the main users' needs. This permitted to initiate the design of prototypes.
The information is organized as follows : notions are grouped in chapters by mathematical domains like geometry, analysis... themself divided in sub-domains, from general to specific notions with videos of signs for each and their definition. The Figure 1 shows an example of the chapter for *Sequences* ("*Suites*" in French, a part of Analysis Domain) with the sub-chapters *Generation*, *Particular sequences*, *Variations* and *Limits*.

In this paper we focus on the question of design of the prototypes' interface and its evaluation. Regarding the visual aspect of the interface, its design is based on the use of tiles, which shape permits a good visualisation of the videos and a tidy organisation of the notions. We had many options to illustrate notions, like textual, diagram, signs in video or static image of sign, and different mixtures of these.

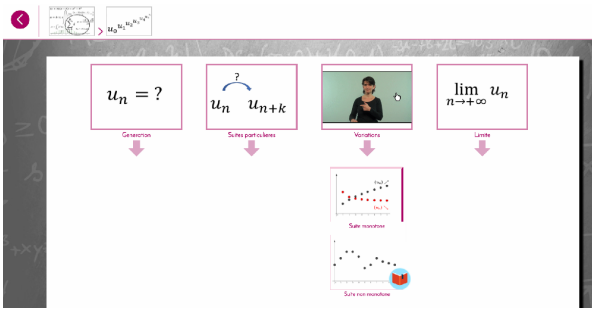


Figure 1: Home page with access via diagrams and text, with an example of SL video launched on mouse-over in the tile for the *Variation* sub-chapter

A first version used diagrams illustrating the notions, textual and video of SL indications – on tiles mouse-over

(Fig. 1). A framed tile represents a chapter and an arrow points at the notions it covers (see *Variations* in the figure). A notion that could be expanded into sub-notions is represented by a stack (for example "*Suite monotone*" - *monotone sequence*). Clicking on the book icon (over "*Suite non monotone*" diagram for instance) open the notion's definition. On the top left of the screen, Ariane's thread and back button allow easy return to previous chapters.

As we had limited time and means for the user evaluation part of the project, we decided to focus the user test part of this work on a radically Sign Language oriented Interface. So in order to investigate if SL alone allow sufficient understanding for the user to efficiently access the definitions, another version relying on SL only was realised (Fig. 2). We chose to make videos play in a loop, an approach supported by Jean-Louis Brugeille (Leroy and Brugeille, 2015), a Deaf expert in SL[1].

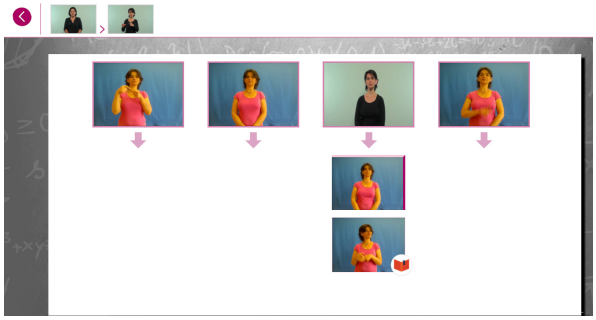


Figure 2: Home page with access via Sign Language

Finally, the definition page is shown in Figure 3. The definition video may include hypermedia links redirecting to prerequisites' definitions when needed. In this example, the prerequisite *Number* is associated with this notion. The moment that it is mentioned (*i.e.* signed) in the definition is represented by a range with white background and blue frame on the progress bar (see the orange circle and arrow). Moreover, the rectangle on the right part of the page shows the indication "Associate notion" and the French translation of the prerequisite. On mouse-over, the rectangle shows *Number* in SL, and on click, the definition page of *Number* opens in a new tab.

[1]Jean-Louis Brugeille has a permanent position in French Ministry for Education as academic inspector in Toulouse for teaching in SL and as national reference for education in SL.

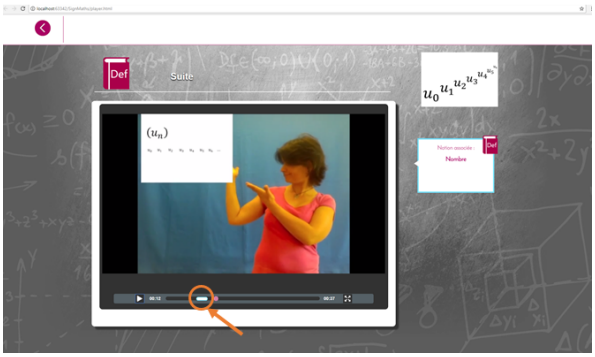


Figure 3: Page definition for *Convergent sequence*

4. User tests on a prototype entirely in Sign Language

The version entirely in SL was put on test with users, which aimed at evaluating if the interface was adapted for the different target audiences: students, teachers and families.

4.1. Participants and goal

The panel of participants we managed to gather was small but matched the user profiles we were looking for: four Deaf high-school students, two bilingual teachers of mathematics in SL, and two Deafs adults. The tests permitted to assess if the education backgrounds of users (*i.e.* education received in SL, French or in both language) triggered significant differences in their performance when using Sign'Maths. At that time the prototype was composed of a few videos (34 mathematical concepts, 17 definitions and 2 examples).

4.2. Test protocol

Very few literature addresses the question of user testing with Deafs. A study conducted by Slegers *et al.* (2010) gives recommendations for involving "target groups with whom researchers and designers cannot communicate as they are used to". However, this study focuses on children. As our participants were older and as we were limited in time, we couldn't follow the same guidelines (*e.g.* spending a whole day with the students at school). More time would have been needed to investigate how to conduct user tests with Deaf teens and adults. Nevertheless, we tried our best to produce a protocol adapted to the situation.

At the beginning of each test session, the participant was asked to fill a pretest questionnaire:

- participant's age: as education for Deafs has significantly changed in the recent years, this information is relevant to understand the background of the participant.
- participant's city of residence and city of his/her studies: as the signs used by Deafs depends on the geographic location, this may bias results.
- if participant followed mathematics lessons on sequences and in which language: this would assess his/her knowledge of the maths vocabulary used during the test.

To reproduce real life cases of use, we asked participants to complete three types of tasks in finding the definition of a mathematical notion. These tasks differs in the way the notion is presented to the participant :

1. signed to him;
2. written in French on a paper;
3. illustrated by a diagram on a paper.

These tasks represent the different situations where students may be searching for a notion's definition in Sign'Maths. These situations are among: they know the sign for the notion (1), they know the French word for the notion (2), or they have encountered a diagram in an exercise (3).
Each task was timed and observations/remarks of the participant were noted down. After each task, the participant was asked to rate the difficulty of the task on a Likert scale (1932) and to explain what changes he/she would make on the interface.
At the end of the test, the participant was asked to fill a SUS questionnaire. Ideally, the SUS questionnaire would have been in SL. However, no official translation existed and our participants did not have sufficient English knowledge - which is understandable as it is their third language - to fill it in the original version. Thus, we chose to use the French version detailed in (Yharrassarry, 2011): even if there is no official French translation yet, this one is the most used of the existing translations.
Finally, we asked the participant if he/she had global remarks on the interface or any propositions of change.

4.3. Setting up and special precautions

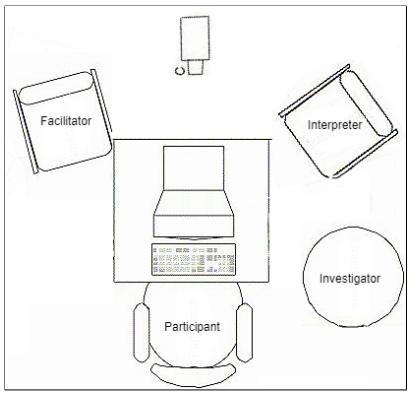


Figure 4: Experimental set-up

The experimental set-up we used for the tests can be seen in the Figure 4. User testing was done on a computer running on Windows 10 on which a stable version of the interface had been set up. The investigator having developed the prototype and having experience in user testing, she was in charge of observing, taking measures and notes, but she had no interaction with the participants during the test.
A person was needed to conduct the test sessions. As most of our participants were Deaf - and all being fluent in SL - this person - the "facilitator" - should communicate in SL. For no other person in our group was familiar with conducting tests, a preliminary training was necessary. This role could not be played by an interpreter as we were not of his/her availability. Thus, a Deaf member of the team was trained to play the facilitator. Work on the different parts of the protocol used has been conducted:

- When asking to open the definition of a signed notion, we chose to provide the sole sign for the notion. If the participant didn't know the sign, we would ask him to find what he/she found the most similar to it.
- For questions on notions in French or illustrated by a diagram, the facilitator gave the user a piece of paper with either the printed word or the printed diagram for the notion. This in order to avoid the misunderstanding of handwritten information and to give all the participants exactly the same information.
- SL being very visual, description of the tasks may include elements of response. For example, when asking for opening a definition, the signer may mime elements of the interface and the interaction needed to open the definition. This visual way to describe an action being intrinsic to SL, there was no other possibility to sign the questions. It is precisely why we had to work on which signs to use to describe elements of the interface. For example, the word "tile" had no translation in SL in this context. The sign which expression was the nearest to the concept of tile was the sign for "rectangle".

To guarantee a fluid communication during the tests, an interpreter SL↔French had been hired. Thus, the investigator could follow the test's progress and write down the participant's comments. Finally, as the interpreter's live translation may not highlight the emotions conveyed by the participant's signs, we chose to record his/her facial expression and body attitude.

4.4. Analysis of the results

Before getting into the results' analysis, it is worth noting that our low number of participants didn't allow to generalise our results. However, it gave us useful insights on the interface tested and allowed us to retrieve design guidelines for the next version.
In this analysis, we compared the performance of 3 categories of users: Deaf students (*Students*), Deaf or bilingual maths teachers (*Teachers*) and Deafs who might use Sign'Maths in the family sphere (*Family*). Their performance was analysed for 3 types of tasks (cf. 4.2.) :

- **Open the definition of a notion signed**:
  The difference in performances is around 10 seconds for the 3 categories of users, *Teachers* being the fastest and *Family* being the slowest, and *Students* being in the middle.
- **Open the definition of a notion written in French**:
  Teachers were 1 minute faster that *Students*, who were 2 minutes faster than *Family*.
- **Open the definition of a notion illustrated by a diagram**:
  Students were 10 seconds faster than *Teachers*, who were 1 minute faster than *Family*.

According to these observations, we can conclude that, when it comes to searching a notion in SL, all users seem to succeed equally. However, users who received education in SL (*i.e.* *Students* and *Teachers*) seem to have fewer difficulties in searching for notions written in French or illustrated by a diagram.
To retrieve conclusions from the SUS scores, we used the article of Bangor *et al.* (2009). The SUS score of *Students* is 61, which is between *OK* and *Good*. The score of *Teachers* is 81 which is between *Good* and *Excellent*. However, the SUS score of *Family* is 33, considered as *Poor*. The global score of our panel of users is 63, meaning that our interface's acceptability falls within the category *Marginal High* - which is a step before the level acceptable.

Finally, the most frequent comment from users was that using the interface required being aware of the mathematics vocabulary used – several signs/words being unknown for some users, especially *Family*. They highlighted that the interface was clear and well organised, and *Students* and *Teachers* particularly appreciated the use of SL. All users agreed that French indications and illustrations of the concepts should be added to the SL content. They also proposed minor changes such as improving videos' quality and adding content to the interface which they found quite simple.

In conclusion, user tests have highlighted that work needed to be done to make our interface more inclusive and more adapted towards users who are not aware of the mathematics vocabulary used in Sign'Maths. Sign'Maths should allow users to find a notion, whether they know its translation in SL, French, or none of them. Eventually, we took advantage of participants' presence to perform a brainstorming session after the test sessions.

5. Final prototype of Sign'Maths

Following the user tests, a low-fidelity prototype has been realised. This prototype is a combination of the two previous versions, and it takes into account the conclusions of the tests and users remarks and propositions.

The chaptering is now materialised by a tree. The main topic is represented by a big tile in the centre (here "*Suites*"). The tiles above are previous chapters; the tiles below are chapters covered by the main one. Each tile:

- is composed of a diagram illustrating the notion;
- shows the notion in SL on mouse-over;
- and is accompanied by the French translation of the notion.

The following figures show the final interface obtained. The tree structure is flexible and automatically adapts to the number of notions displayed. As shown in Figure 5, on tile mouse-over appears the translation of the notion in SL.

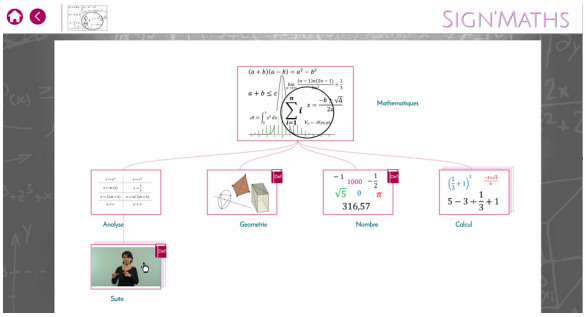


Figure 5: Home page of Sign'Maths

This interface takes advantage of the complementarity of Sign Language, French Language, and visual content. The chaptering is clear and navigation is allowed through several elements (tiles themselves, the Ariane's thread, and the back and home buttons).
Clicking of a tile - in the picture, clicking on the tile "*Suite*" - triggers the *Sequences* chapter expanding (Fig. 6). The tree nodes updates, as well as the Ariane's thread.

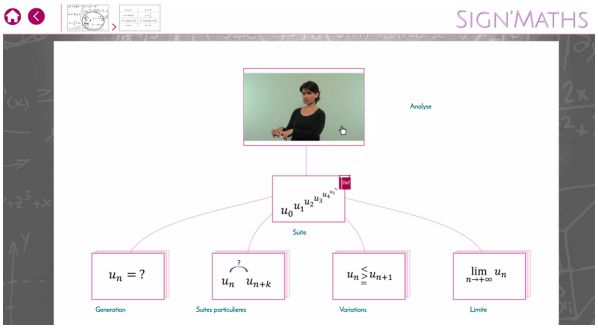


Figure 6: Example of tree exploration in Sign'Maths

Definition and example icons are represented respectively by a purple book with "*Def*" written on the cover and a green book with the indication "*Ex*". A click on these icons triggers the opening of either the definition page (Fig. 3) or the example page - which designs are similar.

6. Conclusions and prospects

Sign'Maths stands out by putting SL at the core of a pedagogical tool. Tests results show that participants appreciate the extensive use of SL in the interface. However, due to the ubiquity of written text in their Education, they agree that French indications are needed to support navigation and would be helpful for a search-by-word tool. Finally, participants appreciate the use of graphics to complement SL and written information. Because mathematics cover abstract concepts, the use of visual information such as diagrams and signs can have a positive impact on Deaf students' learning. The search of unknown mathematical words will be addressed by adding a text search function. Future work will investigate how an entry in SL can be used to search for signs in video definitions.

7. Bibliographical References

Adamo-Villani, N. and Anasingaraju, S., (2017). *Holographic Signing Avatars for Deaf Education*, pages 54–61. Springer International Publishing, Cham.

Bangor, A., Kortum, P., and Miller, J. (2009). Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale. *J. Usability Studies*, 4(3):114–123, May.

Dalle, P. (2003). La place de la langue des signes dans le milieu institutionnel de l'éducation : enjeux, blocages et évolution. *Langue française*, 137(1):32–59.

Debevc, M. and Peljhan, Ž. (2004). The role of video technology in on-line lectures for the deaf. *Disability and Rehabilitation*, 26(17):1048–1059, September.

Jones, M., Bench, N., and Ferons, S. (2014). Vocabulary acquisition for deaf readers using augmented technology. In *VAAT, 2014 2nd Workshop on*, pages 13–15. IEEE.

Khwaldeh, S., Matar, N., and Hunaiti, Z. (2007). Interactivity in Deaf Classroom Using Centralised E-learning System in Jordan. In *The 8^{th} Annual Postgraduate Symposium on the Convergence of Telecommunications, Networking and Broadcasting*, Liverpool.

Kristoffersen, J. and Troelsgård, T., (2012). *The Electronic Lexicographical Treatment of Sign Languages: The Danish Sign Language Dictionary*, pages 293–314. 11.

Leroy, E. and Brugeille, J.-L. (2015). Ressource en autoformation : Langue des Signes Française : quelle évaluation pour une langue sans écriture ? `https://goo.gl/C7fBbM`. Video in French SL, 2015-02.

Likert, R. (1932). A Technique for the Measurement of Attitudes. *Archives of Psychology*, 140:1–55.

McKee, R., (2017). *The Online Dictionary of New Zealand Sign Language: A case study of contemporary sign language lexicography*. 10.

Megat Mohd Zainuddin, N., Badioze Zaman, H., and Ahmad, A., (2009). *Learning Science Using AR Book: A Preliminary Study on Visual Needs of Deaf Learners*, pages 844–855. Springer Berlin Heidelberg, Berlin, Heidelberg.

Slegers, K., Duysburgh, P., and Jacobs, A. (2010). Research Methods for Involving Hearing Impaired Children in IT Innovation. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*, NordiCHI '10, pages 781–784, New York, NY, USA. ACM.

Stokoe, W. (1965). Dictionary of the American Sign Language based on scientific principles.

Straetz, K., Kaibel, A., Raithel, V., Specht, M., Grote, K., and Kramer, F. (2004). An e-learning environment for deaf adults. In *Conference proceedings 8^{th} ERCIM workshop "user interfaces for all"*.

Yharrassarry, R. (2011). Une année d'ergonomie sur le bloc-notes. `http://blocnotes.iergo.fr/tag/sus/`.

Zainuddin, N. M. M., Zaman, H. B., and Ahmad, A. (2010). Developing augmented reality book for deaf in science: the determining factors. In *ITSim, 2010 International Symposium in*, volume 1, pages 1–4. IEEE.

Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 177–180
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

STS-korpus: A Sign Language Web Corpus Tool for Teaching and Public Use

Zrajm Öqvist, Nikolaus Riemer Kankkonen, Johanna Mesch
Department of Linguistics, Stockholm University
SE-106 91 Stockholm, Sweden
{zrajm, nikolaus.kankkonen, johanna.mesch}@ling.su.se

Abstract

In this paper we describe *STS-korpus*, a web corpus tool for Swedish Sign Language (STS) which we have built during the past year, and which is now publicly available on the internet. *STS-korpus* uses the data of Swedish Sign Language Corpus (SSLC) and is primarily intended for teachers and students of sign language. As such it is created to be simple and user-friendly with no download or setup required. The user interface allows for searching – with search results displayed as a simple concordance – and viewing of videos with annotations. Each annotation also provides additional data and links to the corresponding entry in the online Swedish Sign Language Dictionary. We describe the corpus, its appearance and search syntax, as well as more advanced features like access control and dynamic content. Finally we say a word or two about the role we hope it will play in the classroom, and something about the development process and the software used. *STS-korpus* is available here: **https://teckenspraskorpus.su.se**

Keywords: web application, user-focused, accessibility, language teaching, Swedish Sign Language corpora

1. Introduction

Corpora are a valuable tool in second language teaching (Granath, 2009, 245) – however downloading and installing the software and data needed for using a corpus can be a daunting task for students, teachers, and researchers alike.

This is especially relevant in sign language courses for beginners, where no prior knowledge of corpora is expected. These classes would benefit from corpus use, but walking students through a time consuming and complicated installation process takes time away from the primary purpose of the course – language teaching. Thus, there is need for a simpler tool, one which requires less instruction, and has less initial setup, and this is where *STS-korpus* enters the picture (*STS-korpus*, 2020).

For the past year we have been developing *STS-korpus*, a web corpus interface, meaning that we can now simply tell our students to go to **https://teckenspraskorpus.su.se** where they can immediately access our corpora, without the need for them to install anything.

2. Role in Teaching

It is our hope that *STS-korpus* will be used among teachers and students of sign language, both inside and outside of classroom.

STS-korpus was designed specifically with the learning situation in mind, as a simple and fast lookup tool, for use in situations where downloading, installing and configuring a full corpus is beyond the scope of the current exercise, regardless of whether this is because of a lack in technical expertise, available hard drive space, or lack of admin privileges on a classroom computer.

We hope to facilitate the teacher's role by providing easy access to a corpus. This can be used both for making presentations during class, and also as an aid in answering student questions.

The role of the corpus in sign language teaching is important since sign language does not have a written form. Therefore a corpus is the only way to look up how a particular sign is used in everyday conversation.

At the end of the development process we informally reached out to a few teachers of sign language for their opinions. They appreciated the possibilities of the *STS-korpus*, especially the easy availability. This leads us to think that we are on the right track, and that *STS-korpus* can have a future in education.

2.1. In Relation to ELAN

SSLC (Mesch et al., 2012) and later corpora created here at the Department of Linguistics at Stockholm University were all annotated using ELAN (2020). We will therefore here contrast *STS-korpus* with ELAN, rather than iLex (Hanke, 2002) or other tools.

STS-korpus and ELAN have different purposes, and are complementary to each other.

It is only a slight exaggeration to say that ELAN is the Swiss army knife of annotation, a tool that can do almost anything. It can be used for annotation, searching, making statistics and more (Crasborn and Sloetjes, 2010) – everything except real-time collaboration. However it is a complex tool which takes time to learn (Leeson et al., 2019, 345) and for someone who is new to ELAN setting up a corpus for searching and viewing is not a trivial matter.

A course on sign language corpus linguistics would probably not be complete without at least an introduction to ELAN, but for a beginner's course in sign language, introducing a program of ELAN's complexity (Mesch and Wallin, 2008) could be counterproductive.

STS-korpus, on the other hand, is more like an IKEA Allen wrench – designed to do one thing, and to do that with as little fuss as possible. This is far better suited for a beginner's course.

3. Usage

STS-korpus works on both computers and portable devices, though screen size will affect the experience, and a tablet will serve you better than a cell phone.

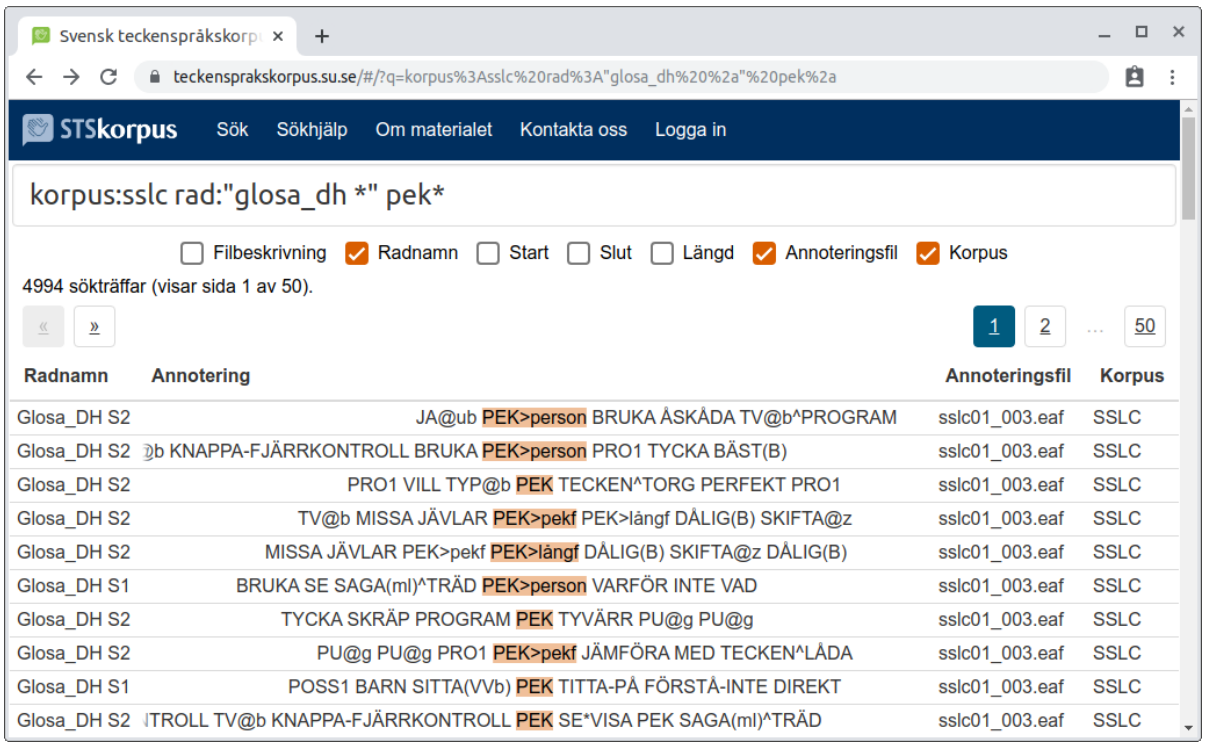


Figure 1: A search with its results

3.1. Searching

STS-korpus has a simple, uncluttered search interface (Figure 1) which is intentionally similar to that of Google and other search engines, so as to feel familiar for both novice and expert computer users.

The user may search the annotations in the database for words, or parts of words (using `*` to match the varying part). Search prefixes may also be used to limit a search to named annotation tiers (`rad:`), files (`fil:`) or corpora (`korpus:`).

Figure 1 shows the result of a search of the Swedish Sign Language Corpus (`korpus:sslc`) in tiers for the dominant hand (`rad:"glosa_dh *"`) for annotations containing pointing (`pek*`).

The tiers available depend on the corpus. SSLC tiers include: `Glosa_DH S1` and `Glosa_DH S2` (glosses for dominant hand of subject 1 and 2), `Glosa_NonDH S1` and `Glosa_NonDH S2` (glosses for the non-dominant hand), and `Översättning S1` and `Översättning S2` (Swedish translations) (cf. Figure 2).

Search results are displayed in a simple concordance view, with each found search term highlighted. The checkboxes can be used to show additional columns of information, such as file names, file descriptions, tier names, corpus names, and start/end times and length of the annotations. (The search results are currently not sorted, but we hope to add options for this in the near future.)

You can also share your result with other people by just copying the web address and sharing that with them. (However, if you are a logged-in registered user, keep in mind that you might have access to information that your recipient will not be able to see [section 4.1.].)

A description of the glossing conventions used in SSLC (Figure 1) can be found in English in Mesch and Wallin (2015), and in Swedish in Wallin and Mesch (to be published 2020).

3.2. Video Viewing

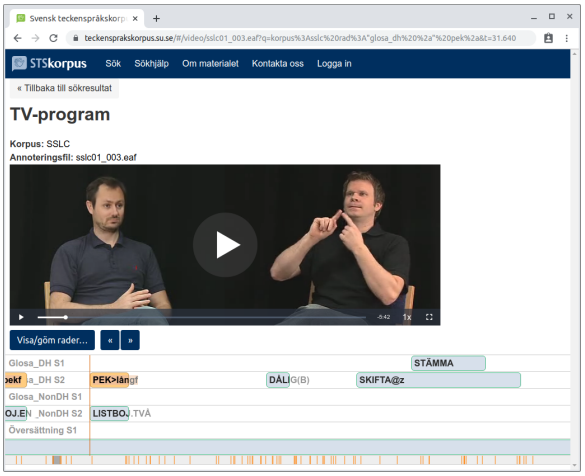


Figure 2: Video view with highlighted search "`PEK*`"

Clicking on any of the matches in the search results will load the video (Figure 2) and skip to the location of the

matching annotation. Matching annotations in the video are highlighted, and the horizontal scrollbar at the bottom of the window has marks to indicate where matches are found in the video, turning the scrollbar into a rough dispersion plot. There are buttons for selecting which annotation tiers to display, and for skipping to the previous and next match. Playback speed can also be adjusted.

This view is similar to ELAN, with annotations scrolling across the screen as the video is playing, and a thin vertical line to indicate the current position. Annotations and video are always kept in sync so that the user may navigate by either scrolling the annotations, or by skipping to a different point using the video player.

3.3. Links to Dictionary

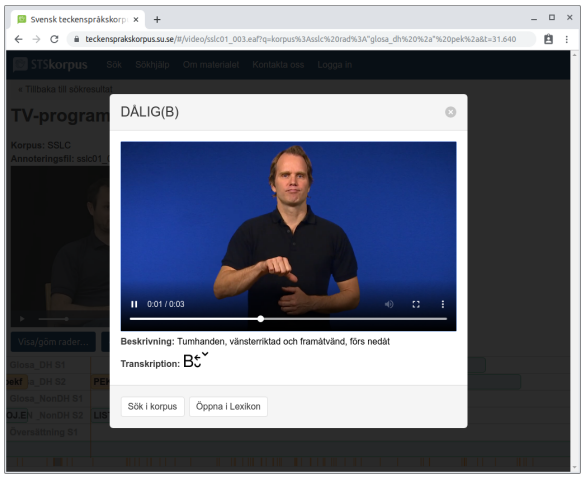


Figure 3: Video view, annotation details

There are links from *STS-korpus* to our already existing online Swedish Sign Language Dictionary (*Svenskt teckenspråkslexikon*, 2008). Clicking on a gloss annotation in the video view quickly brings up additional details about the sign, without having to loading the full dictionary entry (Figure 3). This view includes buttons for performing a new search in the corpus for this particular gloss ("Öppna i korpus") and for going to the full dictionary entry ("Öppna i lexikon").

We have also updated the Swedish Sign Language Dictionary so that each individual entry now contains a link to search for that sign in *STS-korpus*. This facilitates smooth navigation between dictionary and corpus, and makes it very easy to go from a dictionary entry to a list of real-world usage examples for a sign.

3.4. Multiple Datasets

We decided early on to build a backend which would allow for datasets from multiple corpora – a design decision which has served us well. Because of this it has been trivial to add data from other corpora, previously developed at the Department of Linguistics, as interest for the web corpus has grown.

So far we have added data from the Swedish Sign Language Corpus (SSLC) (Mesch et al., 2012), Swedish parts of the ECHO corpus project (Bergman and Mesch, 2004), as well as a special "dynamic corpus" (see section 4.3.). In the future we also hope to add parts of the Tactile Swedish Sign Language Corpus (TSSL) (Mesch, 2016) and the Corpus of Swedish Sign Language as Second Language (SSLC-L2) (Schönström and Mesch, 2017).

4. Advanced Features

4.1. Access Control

Any user can, without having to log in, access the public and anonymized part of SSLC (Mesch et al., 2012). There is also a login system in place, by which we can grant registered users additional access, e.g. access to sensitive or experimental data, as well as data from additional corpora (section 3.4.).

There are three different access levels for registered users: *teacher, researcher* and *admin*.

A *teacher* has permissions suitable for demonstrations in the classroom, i.e. access to a larger number of tiers, but not to sensitive information. A *researcher* has access to everything. And finally an *admin* has access to all data, but can also create and remove users.

4.2. Importing Data

STS-korpus needs two pieces of data for each annotated video: the annotations and the video itself.

All video files in *STS-korpus* were preprocessed, specifically scaled down and reencoded to a format suitable for the web. With the SSLC videos we also took the two camera angles (seen in Figure 2), and merged them into a single video. This was done to avoid possible sync issues while the user is viewing the video which might result from poor browser or computer performance.

Annotations are imported from the ELAN .eaf files used by SSLC, and thereafter put into separate entries in the database.

4.3. Dynamic Content

During early testing teachers at our department requested the ability to upload their own material for use in the classroom. We enabled this by setting up networked hard drive to where teachers may upload their own videos and ELAN annotation files. Files put there are automatically imported into a special corpus, `KURS` (meaning *course*), which makes them available to registered users of *STS-korpus*.

This way registered users, which have also been provided with a separate login for the drive, may upload their files. All uploaded annotations become searchable and viewable in *STS-korpus*, though, for natural reasons, only annotations with glosses that can be found in the Swedish Sign Language Dictionary will link to a dictionary entry.

This way a teacher can upload the data desired for a presentation and then, later on, find it in the web interface by adding `fil:` or `korpus:kurs` to the search.

5. Development

Programming and design of *STS-korpus* was done by Patrick Hansson and Zrajm Öqvist, working one day a week for two semesters.

Throughout the development process we have received continuous feedback from both our colleagues at the office of the Swedish Sign Language Dictionary, and, in the later stages of development, from several other researchers and sign language teachers at the Department of Linguistics.

5.1. Software Used

The frontend was written in *Javascript*, using *Vue.js* and *Buefy*.

The backend was written using *Python*, with the web server and API built using the *Flask* framework. Data is stored in a *MariaDB* database and is served to the frontend by means of a custom made JSON-based API. The database is populated from ELAN .eaf files using our own database import script, written in *Python* and making use of the *SQLAlchemy* and *pympi-ling* modules.

We decided against publishing the source code online, since it is specific to our own setup and was not built with customizability in mind. That being said, we will share the code on request.

6. Conclusion

With this project we have implemented a simple, and easy-to-use sign language corpus for use as a lookup tool in learning situations in the classroom and beyond.

Since *STS-korpus* is intended as a tool mainly for novices and teachers of sign language, our focus has been simplicity of both purpose and design. We have therefore attempted to make *STS-korpus* easy to understand and use. Primarily this means the following:

- Removing the initial hurdle of having to download and install corpus data and ELAN or other similar software.
- Simplifying search and user-friendliness by minimizing clutter. Here we have imitated the appearance of the most commonly used web search engines, and instead of a complex interface with multiple fields and values, we display a single search box. Optional search prefixes (section 3.1.) can be used to perform advanced searches.
- Clearly highlighting the search results in their context. Search results are either shown as a concordance, or as a video with annotations, with search matches clearly highlighted in both cases.

More advanced features include:

- Access control (section 4.1.) by which registered users can be given access to additional corpora or annotation tiers.
- Dynamic content (section 4.3.) which teachers can use to upload their own annotated material for use in the classroom.

7. Acknowledgements

The research reported here was supported in part by the Swedish Sign Language Dictionary at the Department of Linguistics, Stockholm University. None of this would have been possible without the dedicated work of our colleagues Patrick Hansson and Thomas Björkstrand.

8. Bibliographical References

Crasborn, O. and Sloetjes, H. (2010). Using ELAN for Annotating Sign Language Corpora in a Team Setting. In P. Dreuw, et al., editors, *Proceedings of LREC 2010, Fourth Workshop on the Representation and Processing of Sign Languages*, pages 61–64, Valletta, Malta.

ELAN. (2020). [Computer Software]. (Version 5.9) Nijmegen: Max Planck Institute for Psycholinguistics. Retrieved from https://tla.mpi.nl/tools/tla-tools/elan/.

Granath, S. (2009). Who Benefits from Learning How to Use Corpora? In K. Aijmer, editor, *Corpora and Language Teaching*, volume 33 of *Studies in Corpus Linguistics*, pages 47–65. John Benjamin Publishing.

Hanke, T. (2002). iLex – A tool for Sign Language Lexicography and Corpus Analysis. In M. González Rodríguez et al., editors, *Proceedings of LREC 2002, Third International Conference on Language Resources and Evaluation*, pages 923–926.

Leeson, L., Fenlon, J., Mesch, J., Grehan, C., and Sheridan, S. (2019). The Uses of Corpora in L1 and L2/Ln Sign Language Pedagogy. In R.S. Russell, editor, *The Routledge Handbook of Sign Language Pedagogy*, pages 339–352. Routledge, New York.

Mesch, J. and Wallin, L. (2008). Use of Sign Language Materials in Teaching. In *Proceedings of LREC 2008, Sixth International Conference on Language Resources and Evaluation*, pages 134–137.

Mesch, J. and Wallin, L. (2015). Gloss Annotations in the Swedish Sign Language Corpus. *International Journal of Corpus Linguistics*, 20(1):102–120.

Wallin, L. and Mesch, J. (to be published 2020). *Annoteringskonventioner för teckenspråkstexter. Version 8.* [Annotation Conventions for Sign Language Discourse. Version 8]. Sign Language Section, Department of Linguistics, Stockholm University.

9. Language Resource References

Bergman, B. and Mesch, J. (2004). *Dataset. ECHO Dataset for Swedish Sign Language*. Department of Linguistics, Stockholm University.

Mesch, J., Wallin, L., Nilsson, A.-L., and Bergman, B. (2012). *Dataset. Swedish Sign Language Corpus Project 2009–2011 (Version 1)*. Sign Language Section, Department of Linguistics, Stockholm University.

Mesch, J. (2016). *Dataset. Tactile Swedish Sign Language Corpus*. Sign Language Section, Department of Linguistics, Stockholm University.

Schönström, K. and Mesch, J. (2017). *Dataset. Corpus of Swedish Sign Language as Second Language Project 2013–2014 (Version 1)*. Sign Language Section, Department of Linguistics, Stockholm University.

STS-korpus. (2020). [Swedish Sign Language Corpus Web Tool]. Sign Language Section, Department of Linguistics, Stockholm University. Retrieved February 5, 2020, from https://teckensprakskorpus.su.se/.

Svenskt teckenspråkslexikon. (2008). [Swedish Sign Language Dictionary]. Sign Language Section, Department of Linguistics, Stockholm University. Retrieved February 5, 2020, from https://teckenspraklexikon.su.se/.

Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 181–188
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

BosphorusSign22k Sign Language Recognition Dataset

Oğulcan Özdemir[a], Ahmet Alp Kındıroğlu[a], Necati Cihan Camgöz[b], Lale Akarun[a]
[a]Boğaziçi University, Computer Engineering Department, Istanbul, Turkey
[b]CVSSP, University of Surrey, Guildford, United Kingdom
{ogulcan.ozdemir, alp.kindiroglu, akarun}@boun.edu.tr, n.camgoz@surrey.ac.uk

Abstract

Sign Language Recognition is a challenging research domain. It has recently seen several advancements with the increased availability of data. In this paper, we introduce the BosphorusSign**22k**, a publicly available large scale sign language dataset aimed at computer vision, video recognition and deep learning research communities. The primary objective of this dataset is to serve as a new benchmark in Turkish Sign Language Recognition for its vast lexicon, the high number of repetitions by native signers, high recording quality, and the unique syntactic properties of the signs it encompasses. We also provide state-of-the-art human pose estimates to encourage other tasks such as Sign Language Production. We survey other publicly available datasets and expand on how BosphorusSign**22k** can contribute to future research that is being made possible through the widespread availability of similar Sign Language resources. We have conducted extensive experiments and present baseline results to underpin future research on our dataset.

Keywords: Turkish Sign Language (TID), Sign Language Recognition, Deep Learning

1. Introduction

As native languages of the Deaf, Sign Languages (SL) are visio-temporal constructs which convey meaning through hand gestures, upper body motion, facial expressions and mouthings. Automatic Sign Language Recognition (ASLR) is a challenging task and an active research field with the aim of reducing the dependency of sign language interpreters in the daily lives of the Deaf.

Among the many similar problems attempted by deep learning researchers, sign language recognition bears a resemblance to video-based action recognition because of its shared medium of information (Varol et al., 2017), and to speech recognition and machine translation problems (Bahar et al., 2019; Bahdanau et al., 2017), due to its linguistic nature. However, there are certain aspects of ASLR that makes the task more challenging, one of which is the asynchronous multi-articulatory nature of the sign (Sutton-Spence and Woll, 1999). Furthermore, the lack of large databases aimed at computer vision communities and the difficulty of annotating them has been an inhibiting factor in ASLR research (Hanke et al., 2010; Schembri et al., 2013).

In this paper, we present BosphorusSign**22k**, an isolated SL dataset, for benchmarking repeatable deep learning experiments on SLR. The dataset was derived from BosphorusSign (Camgoz et al., 2016c), which has high-quality recordings collected from Deaf users of Turkish Sign Language (TID). The BosphorusSign Dataset was categorized linguistically, where sign glosses with the same meaning but a different set of morphemes were considered belonging to the same class. Although this annotation scheme was aimed to be useful in a Q&A based interaction system, i.e., banking or hospital desk applications (Suzgun et al., 2015), it is not well-suited for sign language recognition and production systems, where distinguishing between instances of similar sign classes with similar manual and non-manual features is essential.

Although, BosphorusSign is publicly available, there are no benchmarks reported on this dataset. Furthermore, BosphorusSign does not have an evaluation protocol, making future research conducted using BosphorusSign dataset incomparable with one another.

Moreover, BosphorusSign dataset only provided skeleton information obtained from the Kinect V2 SDK. Although real-time depth-based skeleton estimation was state-of-the-art at the time of BosphorusSign's creation, it is jittery and lacks the crucial hand pose information, making the skeleton information provided in the BosphorusSign dataset inadequate for training human pose based sign recognition, translation and production models (Stoll et al., 2018; Stoll et al., 2020).

To address these issues, in this paper, we have enhanced and refined the BosphorusSign dataset, to help future research in the fields of sign language recognition and production. The contributions of this work are listed as;

- We have visually reviewed and cleaned up the dataset and removed all erroneous sign performances.
- We have revisited the labeling scheme and converted the linguistic categorization into a form more suitable for recognition and production research where each class shares the same manual features.
- We provide OpenPose (Cao et al., 2018) body and finger coordinates in addition to Kinect V2 skeleton information.
- We have proposed an evaluation protocol and reported two benchmark results using 3D ResNets (Tran et al., 2018) and IDT (Wang and Schmid, 2013) to underpin future research on this dataset.

The rest of this paper is organized as follows: In Section 2., we give an overview of the SLR literature and other publicly available Sign Language Recognition (SLR) datasets. In Section 3., we introduce the new BosphorusSign**22k** dataset. We then describe the baseline methods and share our experimental results in Section 4.. Finally, we conclude this paper in Sections 5. and 6. by analyzing our baseline results, discussion and future work.

2. Related Work

Since the work of Starner et al. (1998), there have been numerous studies on the isolated SLR task. More recently, utilization of state-of-the-art deep learning models (Koller et al., 2019; Zhang et al., 2016) have resulted in better representation learning that is capable of achieving high accuracies over hundreds of unique sign glosses. Because of the spatio-temporal characteristics of the ASLR problem, popular methods from video (human action & activity recognition) (Wang and Schmid, 2013; Carreira and Zisserman, 2017) and speech recognition (Graves and Schmidhuber, 2005) fields have been widely applied with success to the SLR problem. Since the focus of this paper is on computer vision based ASLR methods, the main variations among solutions proposed to this solution lie in their methods of data representation and temporal sign modeling.

Of the present methods in the literature, a large majority use sequences of RGB video frames and/or depth information (Cui et al., 2017; Wang et al., 2016; Wang et al., 2019). Some methods extract additional features from these input sources such as optical flow and coordinates calculated by human pose capture methods such as Kinect (Shotton et al., 2012) and OpenPose (Cao et al., 2018). A large number of methods extract state of the art features such as 2D (Koller et al., 2016a) and 3D Convolutional Neural Network (CNN) outputs (Joze and Koller, 2018; Huang et al., 2015; Camgoz et al., 2016a), Improved Dense Trajectory (IDT) features (Özdemir et al., 2016), hand appearance and trajectory features (Özdemir et al., 2018; He et al., 2016; Metaxas et al., 2018). The use of spatial attention as in Yuan et al. (2019) to focus learning on the signing space and temporal attention as in Camgoz et al. (2017) and Camgoz et al. (2018; Camgoz et al. (2020) are also among current popular research directions.

In terms of temporal segmentation and modeling, isolated videos of sign glosses often consist of varying length and complexity, requiring the use of temporal modeling. In Aran (2008) and Zhang et al. (2016), Hidden Markov Models (HMMs) are used with hand shape features and trajectories, while Koller et al. (2016b) uses HMMs to train hand shape classifiers from weakly labeled sign videos. In Liu et al. (2016), Long Short-Term Memory Networks (LSTMs) are used with gradient histograms while in Camgoz et al. (2017) they are used with Connectionist Temporal Classification (CTC) and neural network features to learn sign languages. Based on the works of Joze and Koller (2018) and Camgoz et al. (2016a) and the results of popular gesture recognition challenges such as Chalearn LAP (Wan et al., 2017), SLR studies with 3D CNNs currently tend to show higher performances in large datasets compared to other deep learning approaches utilizing LSTMs and other popular methods. This generalization loses validity in the case of continuous SLR where the average clip length exceeds a few seconds.

In isolated Sign Language Recognition, the difficulties in obtaining high quality annotated videos has limited the amount of available public datasets. To the best of our knowledge, currently there exist four similar public available large scale isolated SLR datasets, which can be seen in Table 1, while Chinese Sign Language (CSL) recognition dataset (Zhang et al., 2016) and MS-ASL (Joze and Koller, 2018) being the most recent ones. BosphorusSign**22k** differentiates from these datasets as follows: Contrary to Chinese Sign Language (CSL) recognition dataset, which was recorded in front of a white background, BosphorusSign**22k** dataset was captured using a Chroma Key background, which we believe will be beneficial for researchers who would like to utilize data augmentation techniques in their pipelines to improve their models generalization capabilities. We acknowledge the fact that MS-ASL is one of the new frontiers in sign language research as it initiated large scale "in the wild" isolated sign language recognition. However, the dataset is composed of publicly available YouTube videos. As Microsoft does not own the copyright of these videos, the availability of the data is not guaranteed. As of the submission of this paper, 290 videos are no longer publicly available which will potentially increase in the future, making comparison of future research against previously reported benchmarks harder. Additionally, there is the new SMILE DSGS corpus (Ebling et al., 2018), which hasn't been fully publicly available and its evaluation protocol is yet to be defined.

3. BosphorusSign22k Dataset

In this study, we present BosphorusSign**22k**[1], a new benchmark dataset for vision-based user-independent isolated SLR. Our dataset is based on the BosphorusSign (Camgoz et al., 2016c) corpus which was collected with the purpose of helping both linguistic and computer science communities. It contains isolated videos of Turkish Sign Language glosses from three different domains: Health, finance and commonly used everyday signs. Videos in this dataset were performed by six native signers, as shown in Figure 1, which makes this dataset valuable for user independent sign language studies.

All of the sign video recordings in the dataset were captured using Microsoft Kinect v2 (Zhang, 2012) with 1080p (1920x1080 pixels) video resolution at 30 frames per second. We believe having a higher resolution is essential for sign language recognition when interpreting the appearance information related to hand shape and movements. All of the videos share the same recording setup where signers stood in front of a Chroma-Key background which is 1.5 meter far away from the camera.

Specifications of the BosphorusSign**22k** dataset can be seen in Table 2. Since the dataset was collected using Microsoft Kinect v2, we provide RGB video, depth map and skeleton information of the signer for all sign videos in the dataset. Moreover, we also provide OpenPose (Cao et al., 2018) joints, which include facial landmarks and hand joint positions in addition to body pose information. An example of provided modalities of the BosphorusSign**22k** dataset can be seen in Figure 2.

BosphorusSign**22k** has a vocabulary of 744 sign glosses; 428 in Health while having 163 in Finance domains as well as another 174 commonly used sign glosses. Properties of the proposed dataset and how it differentiates from the BosphorusSign corpus can be found in Table 3.

[1] https://www.bosphorussign.com

Table 1: Publicly available Isolated Sign Language Recognition datasets

Dataset	Sign Language	#Signers	Lexicon	Repetitions	#Clips	All Native Signers	Data Source
ASLLVD (Neidle et al., 2012)	American	6	2,742	arbitrary	9,794	Yes	RGB
Devisign (Chai et al., 2014)	Chinese	8	2,000	1-2	24,000	No	Kinect v1
BosphorusSign (Camgoz et al., 2016c)	Turkish	6	855	4+	22,670	Yes	Kinect v2
CSL (Zhang et al., 2016)	Chinese	50	500	5	125,000	No	Kinect v2
MS-ASL (Joze and Koller, 2018)	American	222	1,000	arbitrary	25,513	Yes	RGB
BosphorusSign22k	**Turkish**	**6**	**744**	**4+**	**22,542**	**Yes**	**Kinect v2**

Figure 1: Native signer participants of the BosphorusSign**22k** dataset.
(We propose using the left-most five signers as the training set and keep the remaining for evaluation.)

Table 2: Specifications of the BosphorusSign**22k** dataset.

Property	Description
Number of sign classes	744
Number of signers	6
Number of videos	22,542
Total Duration	∼19 hours (∼2M frames)
RGB Resolution	1920 x 1080 pixels
Depth Resolution	512 x 424 pixels
Frame Rate	30 frames/second
Body Pose Information (Kinect v2)	25 x 3D Keypoints
Body Pose Information (OpenPose)	25 x 2D Keypoints
Facial Landmarks (OpenPose)	70 x 2D Keypoints
2 x Hand Pose Information (OpenPose)	21 x 2D Keypoints

Table 3: Properties of the publicly available subsets of the BosphorusSign corpus and the proposed BosphorusSign**22k** datasets.

Dataset	Lexicon	# Clips	# Repetitions
HospiSign (Camgoz et al., 2016b)	33	1,257	6-8
BosphorusSign (Camgoz et al., 2016c)	855	–	–
- Publicly Available	595	22,670	4+
BosphorusSign22k	744	22,542	4+
- General	174	5,788	4+
- Finance	163	4,998	4+
- Health	428	11,756	4+

In this work, we changed several aspects of the BosphorusSign dataset. First of all to set a baseline that would extend over the whole dataset, we merged all subsets and conducted our experiments accordingly. Further details of our experimental protocol can be found in the Section 4.1. Secondly, we manually inspected all the sign videos and eliminated erroneous recordings. Furthermore, we split signs that were semantically same but morphologically different. We also collapsed signs with similar manual features. The goal of this change was to benchmark the capabilities of the state-of-the-art models on learning meaningful representations for manual aspects of the sign glosses. However, we will be also releasing an uncollapsed version of the dataset. The changes on the BosphorusSign dataset are mostly focused on improving and cleaning the dataset and defining an evaluation protocol. The dataset will be publicly available for research purposes upon submitting an EULA to the authors.

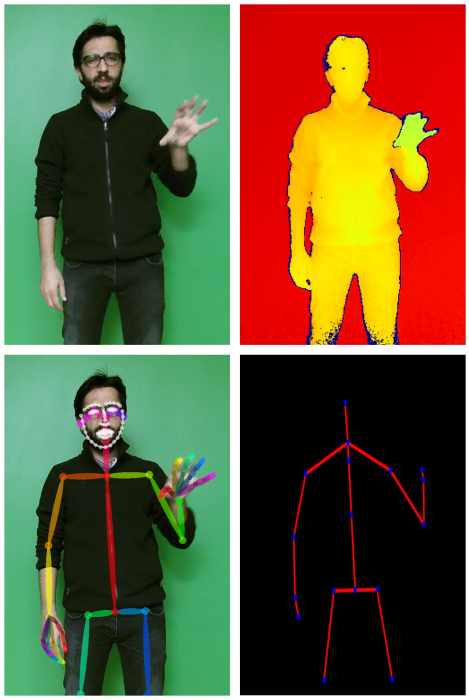


Figure 2: Modalities of BosphorusSign**22k**.
Top: (left) RGB frame and (right) depth image
Bottom: (left) OpenPose and (right) Kinect v2 outputs.

4. Baseline Recognition Methods and Experiments

In this section, we provide training and test protocols of the BosphorusSign**22k** dataset. We then describe baseline methods in detail and share our extensive experimental results.

4.1. Experiment Protocol

We defined our protocol by dividing the BosphorusSign**22k** dataset into training and test sets in a signer independent manner where we use one signer for test and others for training. This yielded us a test set of 4,524 sign samples and a training set containing 18,018 samples. To set a baseline on the new dataset and the evaluation protocols, we perform isolated sign language recognition experiments and report classification accuracies on the test set with only one signer.

4.2. Baseline Methods

As for benchmark methods, there is no agreed-upon state-of-the-art approach in the isolated SLR literature, and researchers use different methods on different datasets (Camgoz et al., 2017; Joze and Koller, 2018; Zhang et al., 2016). However, in the related field of action recognition, researches rely on benchmark datasets to compare their approaches against the state-of-the-art. Both 3D ResNets (Tran et al., 2018) and Improved Dense Trajectories - IDT (Wang and Schmid, 2013) are comparable state-of-the-art methods for action recognition, which have also yielded good performance on SLR (Özdemir et al., 2016). Therefore, we have chosen 3D ResNets and IDT as our baseline approaches to cover both deep learning based representation learning techniques as well as hand crafted feature based methods.

Improved Dense Trajectories - IDT: Although deep learning based models have become very popular recently, handcrafted approaches are still representative and competitive enough to be used in video recognition problems such as human action recognition and sign language recognition (Tran et al., 2018). Inspired by this and to also give further insight to the reader instead of just reporting baselines using a deep learning based approach, we have used Improved Dense Trajectories (IDT) (Wang and Schmid, 2013) which is one of the most successful handcrafted methods with competitive performance for human action recognition and was used in sign language and gesture recognition recently (Özdemir et al., 2016; Peng et al., 2015). IDT extracts local spatial features Histograms of Oriented Gradients (HOG) (Dalal and Triggs, 2005), and local temporal features Histograms of Optical Flow (HOF) (Laptev et al., 2008) and Motion Boundary Histograms (MBH) (Dalal et al., 2006) from the trajectories computed from dense optical flow field.

For recognizing sign language videos from the BosphorusSign**22k** dataset, we have used a recognition pipeline similar to the one proposed by Wang and Schmid (2013). We first extracted trajectories from every sign video. After extracting trajectories, we randomly sampled trajectories from the training set, assuming these trajectories represent the overarching distribution. Then, Principal Component Analysis (PCA) was applied to each component; namely HOG, HOF and MBH. Using PCA outputs, we performed Gaussian Mixture Model (GMM) to cluster each component of the trajectories. Finally, Fisher Vectors (FVs) were computed from each component of trajectory descriptors from each sign video using the parameters of PCA and GMM. Using these representations we trained a Linear Support Vector Machines (SVMs) using different combinations of concatenated FVs components.

3D Residual Networks with Mixed Convolutions: With the recent success of deep learning (Goodfellow et al., 2016) on tasks such as image and object recognition (Krizhevsky et al., 2012), researchers have also started to build deep architectures for human action recognition where both spatial and temporal dimension are exploited (Simonyan and Zisserman, 2014; Tran et al., 2015; Carreira and Zisserman, 2017). In this work, we have used a recently proposed video recognition method, which is based on 3D ResNet architecture with mixed 2D-3D convolutions, also called MC3 in Tran et al. (2018). The model consist of two residual blocks with 3D convolutions, three residual blocks with 2D convolutions and a fully connected layer as its classification layer. This network was built based on the hypothesis that learning temporal dynamics is beneficial in early layers while the higher levels semantic knowledge can be learnt in late layers (Tran et al., 2018).

As our second baseline, we trained MC3 models and investigated the effects of fine-tuning different residual blocks of the network using the Kinetics-400 dataset (Carreira and Zisserman, 2017). The proposed training method for this model used randomized clips of frames as inputs, which is not suitable for our problem because randomized clips may include different non-recurring parts with the same isolated sign gloss. Therefore, at the training phase, we randomly sampled batches of uniformly sampled frames from sign videos to give our networks sufficient coverage over the frames.

4.3. Implementation Details

To evaluate baseline methods on the BosphorusSign**22k** dataset, we have used the publicly available implementation in Wang and Schmid (2013) for extracting IDT features and PyTorch (Paszke et al., 2017) implementation of 3D ResNet model with mixed convolutions.

During training, we uniformly sampled 16 frames form each sign gloss with sizes of 112x112 pixels before feeding them to our networks. Sign clips are horizontally flipped a probability of 0.5 to be able to generalize over signers with different dominant hands. After preprocessing, we train our network using Adam optimizer (Kingma and Ba, 2014) with batch size of 32 on a NVIDIA Tesla V100 GPU. For testing, we performed the same preprocessing approach as in training except horizontal flipping of frame clips.

4.4. Experimental Results

We start our experiments by training several IDT based model with varying feature components, results of which can be seen in Table 4. Our results have shown that using motion features separately, HOF and MBH, has yielded better results (83.29% and 86.63%) than using only appearance features, trajectory information (63.68%) and HOG

Table 4: Baseline IDT results on the BosphorusSign**22k** dataset

Method	Top-1 Acc (%)
TRAJ	63.68
HOG	76.59
HOF	83.29
MBH	86.63
HOG + HOF	86.89
HOG + MBH	86.98
HOF + MBH	87.33
HOG + HOF + MBH	**88.53**
TRAJ + HOG + HOF + MBH	87.86

(76.59%). Since the trajectory information only contains the position of the trajectory (mostly the position of hand regions in our case), it is expected that it cannot fully represent motion or appearance based features which have higher dimensionality and cannot encode more complex information about the sign or hand shape.

Moreover, in the case of fusion of the trajectory components, our experiments have shown that using only HOG, HOF and MBH features together improves our recognition accuracy (88.53%), while adding the trajectory (TRAJ) information sightly decreases the performance of our system (87.86%). Although the performance of our system is very close in the case of fusing multiple components, we can see that using HOG features with other features has improved our recognition performance in all cases. This supports the idea that appearance representation obtained using hand crafted features, such as HOG, is useful along with the temporal information when recognizing signs where the manual features of the sign are the main differentiating aspect between target classes.

As for our deep learning baseline, we performed several experiments on fine-tuning different residual blocks of the MC3 network (Tran et al., 2018). In our first experiment, we first compared training our networks from scratch against using a pre-trained network (on Kinetics-400 dataset) and only training the final fully connected (fc) layer.

Table 5: Baseline 3D ResNet results on the BosphorusSign**22k** dataset

Method	Top-1 Acc (%)	Top-5 Acc (%)
Training from scratch	57.76	84.22
Training only the last fc	55.03	81.98
Fine-tuning last 2 blocks	75.38	94.16
Fine-tuning last 3 blocks	**78.85**	**94.76**
Fine-tuning last 4 blocks	63.88	88.66
Fine-tuning all layers	71.02	92.51

As it can be seen in Table 5, training the network from scratch performed slightly better. We believe this is due to the fact that pre-trained network has never seen any sign samples, hence some of the essential spatio-temporal information that forms the sign is lost until it reaches the final fully connected layer. Using this insight, we decided to fine-tune other layers of the pre-trained network in additional to the last fully connected layer. We found that fine tuning the last 3 blocks to yield the best results for our task, an Top-1 accuracy of 78.85% and and a Top-5 accuracy of 94.76% on the test set.

5. Analysis of Results and Discussion

Although the 88.53% Top-1 accuracy achieved by IDT is quite high for a 744 class problem, there is certainly still room for improvement. On the other hand, 3D ResNets, which are general-purpose video classification algorithms perform worse. We believe this is due to their inability to model longer-range temporal characteristics. Possible improvements include additional modalities and better temporal modeling.

We further investigated false classifications to gain further insight. For example, INSURANCE, INTERNET and COLD sign glosses are commonly confused with FUND, TEACHING and FACE respectively. Upon investigation we discovered that this is due to baseline methods' inability to model fine grained hand shapes. As seen in Figure 3, while our models were able to distinguish the sign by its motion, it fails to discriminate it using the similar handshapes. We believe this is due to the image resolution our networks are trained for in the case of MC3 model and the representation limitations of the HOG features for our IDT baseline. One way to tackle this problem could be to utilize specialized networks, such as Deep Hand proposed by Koller et al. (2016a), and use it as another modality in our recognition pipeline.

Figure 3: Test sample of INSURANCE sign gloss (left) misclassified as FUND (right).

In addition, our analysis have shown that PRICE sign gloss is confused with SHOPPING sign gloss (see Figure 4) because the number of repetitions of the same motion sub-unit is different in both signs. Although both signs have the same hand shape and movements, signers performing the SHOPPING sign gloss repeat the same motion sub-unit more than PRICE.

Furthermore, when looking at the Top-5 recognition accuracy on experiments with 3D ResNets, we can see that most of the misclassified signs are successfully classified among the top-5. Thus, we believe that focusing on problems mentioned above will help us to improve recognition performance. Baseline results also show that IDT, as a handcrafted approach, is still performing well on SLR as it can comprehensively model appearance and motion information obtained from the signer in the frame compared

Figure 4: Test sample of PRICE sign (left) misclassified as SHOPPING (right)

to the 3D ResNet model trained without any specific guidance. Another factor contributing to this performance disparity is the input size which is 112x112 for MC3 networks and 640x360 for the IDT.

6. Conclusion

In this paper, we present BosphorusSign**22k**, a new signer-independent SLR evaluation benchmark. The dataset contains over 22k samples of isolated videos, of 744 unique Turkish Sign glosses performed by six native signers. To underpin future research, we applied two successful video recognition methods from the literature, namely IDT and 3D ResNets (MC3). We share our quantitative results as well as qualitative samples, providing further insight to the reader.

As shown by our experimental results, there is still room for improvement in signer-independent SLR for cases where the manual aspects of the sign subtlety differentiates from other classes. As future work we plan to exploit the capture setup of our dataset, namely its suitability for data augmentation, and extend our protocol to investigate environment independent SLR. BosphorusSign**22k** also enables further research into using the depth information to explore multi-modal fusion approaches.

7. Acknowledgements

This work has been supported by the TUBITAK Project #117E059 and TAM Project #2007K120610 under the Turkish Ministry of Development.

8. Bibliographical References

Aran, O. (2008). *Vision-based Sign Language Recognition: Modeling and Recognizing Isolated Signs with Manual and Non-Manual Components*. Ph.D. thesis, Bogazici University.

Bahar, P., Bieschke, T., and Ney, H. (2019). A comparative study on end-to-end speech to text translation. *arXiv preprint arXiv:1911.08870*.

Bahdanau, D., Bosc, T., Jastrzębski, S., Grefenstette, E., Vincent, P., and Bengio, Y. (2017). Learning to Compute Word Embeddings on the Fly. *arXiv:1706.00286*.

Camgoz, N. C., Hadfield, S., Koller, O., and Bowden, R. (2016a). Using convolutional 3d neural networks for user-independent continuous gesture recognition. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 49–54. IEEE.

Camgoz, N. C., Kindiroglu, A. A., and Akarun, L. (2016b). Sign Language Recognition for Assisting the Deaf in Hospitals. In *Proceedings of the International Workshop on Human Behavior Understanding (HBU)*.

Camgoz, N. C., Kindiroglu, A. A., Karabuklu, S., Kelepir, M., Ozsoy, A. S., and Akarun, L. (2016c). BosphorusSign: A Turkish Sign Language Recognition Corpus in Health and Finance Domains. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Camgoz, N. C., Hadfield, S., Koller, O., and Bowden, R. (2017). Subunets: End-to-end hand shape and continuous sign language recognition. In *IEEE International Conference on Computer Vision (ICCV)*.

Camgoz, N. C., Hadfield, S., Koller, O., Bowden, R., and Ney, H. (2018). Neural Sign Language Translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Camgoz, N. C., Koller, O., Hadfield, S., and Bowden, R. (2020). Sign Language Transformers: Joint End-to-end Sign Language Recognition and Translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., and Sheikh, Y. (2018). OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*.

Carreira, J. and Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, July.

Chai, X., Wang, H., and Chen, X. (2014). The devisign large vocabulary of chinese sign language database and baseline evaluations. *Technical report VIPL-TR-14-SLR-001. Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS*.

Cui, R., Liu, H., and Zhang, C. (2017). Recurrent Convolutional Neural Networks for Continuous Sign Language Recognition by Staged Optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. IEEE.

Dalal, N., Triggs, B., and Schmid, C. (2006). Human detection using oriented histograms of flow and appearance. In *Computer Vision – ECCV 2006*, pages 428–441. Springer Berlin Heidelberg.

Ebling, S., Camgoz, N. C., Braem, P. B., Tissi, K., Sidler-Miserez, S., Stoll, S., Hadfield, S., Haug, T., Bowden, R., Tornay, S., Razavi, M., and Magimai-Doss, M. (2018). SMILE Swiss German Sign Language Dataset. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT press.

Graves, A. and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610.

Hanke, T., König, L., Wagner, S., and Matthes, S. (2010). DGS Corpus & Dicta-Sign: The Hamburg Studio Setup. In *Proceedings of the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*.

He, J., Liu, Z., and Zhang, J. (2016). Chinese sign language recognition based on trajectory and hand shape features. In *2016 Visual Communications and Image Processing (VCIP)*, pages 1–4. IEEE.

Huang, J., Zhou, W., Li, H., and Li, W. (2015). Sign language recognition using 3d convolutional neural networks. In *2015 IEEE international conference on multimedia and expo (ICME)*, pages 1–6. IEEE.

Joze, H. R. V. and Koller, O. (2018). MS-ASL: A large-scale data set and benchmark for understanding american sign language. *CoRR*, abs/1812.01053.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Koller, O., Ney, H., and Bowden, R. (2016a). Deep Hand: How to Train a CNN on 1 Million Hand Images When Your Data Is Continuous and Weakly Labelled. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Koller, O., Zargaran, S., Ney, H., and Bowden, R. (2016b). Deep Sign: Hybrid CNN-HMM for Continuous Sign Language Recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*.

Koller, O., Camgoz, N. C., Bowden, R., and Ney, H. (2019). Weakly Supervised Learning with Multi-Stream CNN-LSTM-HMMs to Discover Sequential Parallelism in Sign Language Videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*.

Laptev, I., Marszalek, M., Schmid, C., and Rozenfeld, B. (2008). Learning realistic human actions from movies. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE.

Liu, T., Zhou, W., and Li, H. (2016). Sign language recognition with long short-term memory. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 2871–2875. IEEE.

Metaxas, D., Dilsizian, M., and Neidle, C. (2018). Linguistically-driven framework for computationally efficient and scalable sign recognition. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Neidle, C., Thangali, A., and Sclaroff, S. (2012). Challenges in development of the american sign language lexicon video dataset (asllvd) corpus. In *5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, LREC*. Citeseer.

Özdemir, O., Camgöz, N. C., and Akarun, L. (2016). Isolated sign language recognition using improved dense trajectories. In *2016 24th Signal Processing and Communication Application Conference (SIU)*, pages 1961–1964. IEEE.

Özdemir, O., Kindiroglu, A. A., and Akarun, L. (2018). Isolated sign language recognition with fast hand descriptors. In *2018 26th Signal Processing and Communications Applications Conference (SIU)*, pages 1–4. IEEE.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*.

Peng, X., Wang, L., Cai, Z., and Qiao, Y. (2015). Action and gesture temporal spotting with super vector representation. In *Computer Vision - ECCV 2014 Workshops*, pages 518–527. Springer International Publishing.

Schembri, A., Fenlon, J., Rentelis, R., Reynolds, S., and Cormier, K. (2013). Building the British Sign Language Corpus. *Language Documentation & Conservation (LD&C)*, 7.

Shotton, J., Girshick, R., Fitzgibbon, A., Sharp, T., Cook, M., Finocchio, M., Moore, R., Kohli, P., Criminisi, A., Kipman, A., et al. (2012). Efficient human pose estimation from single depth images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(12):2821–2840.

Simonyan, K. and Zisserman, A. (2014). Two-stream Convolutional Networks for Action Recognition in Videos. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*.

Starner, T., Weaver, J., and Pentland, A. (1998). Real-time american sign language recognition using desk and wearable computer based video. *IEEE Transactions on pattern analysis and machine intelligence*, 20(12):1371–1375.

Stoll, S., Camgoz, N. C., Hadfield, S., and Bowden, R. (2018). Sign Language Production using Neural Machine Translation and Generative Adversarial Networks. In *Proceedings of the British Machine Vision Conference (BMVC)*.

Stoll, S., Camgoz, N. C., Hadfield, S., and Bowden, R. (2020). Text2Sign: Towards Sign Language Production using Neural Machine Translation and Generative Adversarial Networks. *International Journal of Computer Vision (IJCV)*.

Sutton-Spence, R. and Woll, B. (1999). *The Linguistics of British Sign Language: An Introduction*. Cambridge University Press.

Suzgun, M. M., Ozdemir, H., Camgoz, N. C., Kindiroglu, A., Basaran, D., Togay, C., and Akarun, L. (2015). Hospisign: An Interactive Sign Language Platform for Hearing Impaired. In *Proceedings of the International Conference on Computer Graphics, Animation and Gaming Technologies (Eurasia Graphics)*.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2015). Learning Spatiotemporal Features with 3D Convolutional Networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., and Paluri, M. (2018). A closer look at spatiotemporal convolutions for action recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6450–6459.

Varol, G., Laptev, I., and Schmid, C. (2017). Long-term temporal convolutions for action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1510–1517.

Wan, J., Escalera, S., Anbarjafari, G., Jair Escalante, H., Baró, X., Guyon, I., Madadi, M., Allik, J., Gorbova, J., Lin, C., et al. (2017). Results and analysis of chalearn lap multi-modal isolated and continuous gesture recognition, and real versus fake expressed emotions challenges. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3189–3197.

Wang, H. and Schmid, C. (2013). Action recognition with improved trajectories. In *2013 IEEE International Conference on Computer Vision*, pages 3551–3558. IEEE.

Wang, H., Chai, X., Hong, X., Zhao, G., and Chen, X. (2016). Isolated Sign Language Recognition with Grassmann Covariance Matrices. *ACM Transactions on Accessible Computing*, 8(4).

Wang, H., Chai, X., and Chen, X. (2019). A novel sign language recognition framework using hierarchical grassmann covariance matrix. *IEEE Transactions on Multimedia*, 21(11):2806–2814.

Yuan, Q., Wan, J., Lin, C., Li, Y., Miao, Q., Li, S. Z., Wang, L., and Lu, Y. (2019). Global and local spatial-attention network for isolated gesture recognition. In *Chinese Conference on Biometric Recognition*, pages 84–93. Springer.

Zhang, J., Zhou, W., Xie, C., Pu, J., and Li, H. (2016). Chinese sign language recognition with adaptive hmm. In *2016 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Zhang, Z. (2012). Microsoft kinect sensor and its effect. *IEEE MultiMedia*, 19(2):4–10, April.

Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 189–196
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

Unsupervised Term Discovery for Continuous Sign Language

Korhan Polat, Murat Saraçlar
Electrical & Electronics Engineering Department
Bogazici University, Istanbul, Turkey
{korhan.polat, murat.saraclar}@boun.edu.tr

Abstract

Most of the sign language recognition (SLR) systems rely on supervision for training and available annotated sign language resources are scarce due to the difficulties of manual labeling. Unsupervised discovery of lexical units would facilitate the annotation process and thus lead to better SLR systems. Inspired by the unsupervised spoken term discovery in speech processing field, we investigate whether a similar approach can be applied in sign language to discover repeating lexical units. We adapt an algorithm that is designed for spoken term discovery by using hand shape and pose features instead of speech features. The experiments are run on a large scale continuous sign corpus and the performance is evaluated using gloss level annotations. This work introduces a new task for sign language processing that has not been addressed before.

Keywords: unsupervised learning, term discovery, sign language recognition

1. Introduction

Despite the recent advancements in computer vision and deep learning, automatic sign language recognition (ASLR) still remains as a challenging problem and has the potential for improvement. One of the many reasons that hinders development of ASLR systems is the lack of large scale annotated corpora for training supervised deep learning models. Even though there exist plenty of sign language recordings, most of them are not annotated because manual annotation is a labor intensive task which requires linguistic expertise. This brings the need for a language independent, unsupervised learning procedure in order to handle the vast amount resources for sign languages. With this target set, we explore how an unsupervised learning technique in speech processing can be applied in sign language domain to identify lexical structures when there is no labeled data available.

Unsupervised learning has been an active research area in spoken language processing since the majority of the world's languages are low resource in the sense that there are not adequate resources for training models. The extreme case for unsupervised learning, in which there is neither labeled training data nor knowledge of linguistic structure, is referred as the *zero resource* setting (Versteegh et al., 2015; Dunbar et al., 2017). Zero resource speech processing research focuses on two main topics; subword modelling and spoken term discovery. Subword modelling aims to learn speech representations that capture linguistic structures and that are robust for speech recognition. On the other hand, the aim of unsupervised term discovery (UTD) is to find repeating patterns (phonological, lexical or phrasal units) given only the speech features extracted from raw acoustic signals, without any supervision. The output is the hypothesized word types together with token time boundaries for the unknown language. The pioneering work in unsupervised spoken term discovery by Park and Glass (2008) introduces the segmental dynamic time warping (sDTW) algorithm to discover similar segments between two vector time series. Discovered segments are then clustered where each cluster represents the hypothesized word type in that unknown language. Follow up work of Jansen and Durme (2011) proposes an algorithm that reduces the time complexity by applying efficient image processing and randomized bit hashing techniques. Since then, various approaches to this problem have been proposed in Zero Resource Speech Challenges (Versteegh et al., 2015; Dunbar et al., 2017), which are not in the scope of this work.

Here, we define a new task for processing of sign language videos. Unsupervised discovery of sign terms is the task of discovering and segmenting sign glosses automatically, without using any supervisory information (additional modalities, lexical knowledge etc.). This task would provide numerous benefits to sign language and action recognition fields. It can be used as a segmentation tool that proposes gloss time boundaries and it can speed up manual annotation process. Moreover, clustered segments can be treated as weak labels and supervised models can be trained based on these labels. As an initial exploration of this task, we use the method of Jansen and Durme (2011) since it has been used as the baseline method for the ZR Challenges (Versteegh et al., 2015; Dunbar et al., 2017) and its software implementation is publicly available[1]. We adapt this algorithm to run with sign language videos by feeding visual features instead of speech features. Visual features include hand shape and pose features obtained from pre-trained models. We further augment the pose features by training an autoencoder, which is also an unsupervised learning method. The discovery algorithm is run with these features on a large scale continuous sign dataset and results are evaluated using a set of metrics tailored for this task.

In the field of unsupervised sign language recognition, a similar work to ours is presented by Nayak et al. (2012). They propose a Bayesian method to find the most oc-

[1] github.com/arenjansen/ZRTools

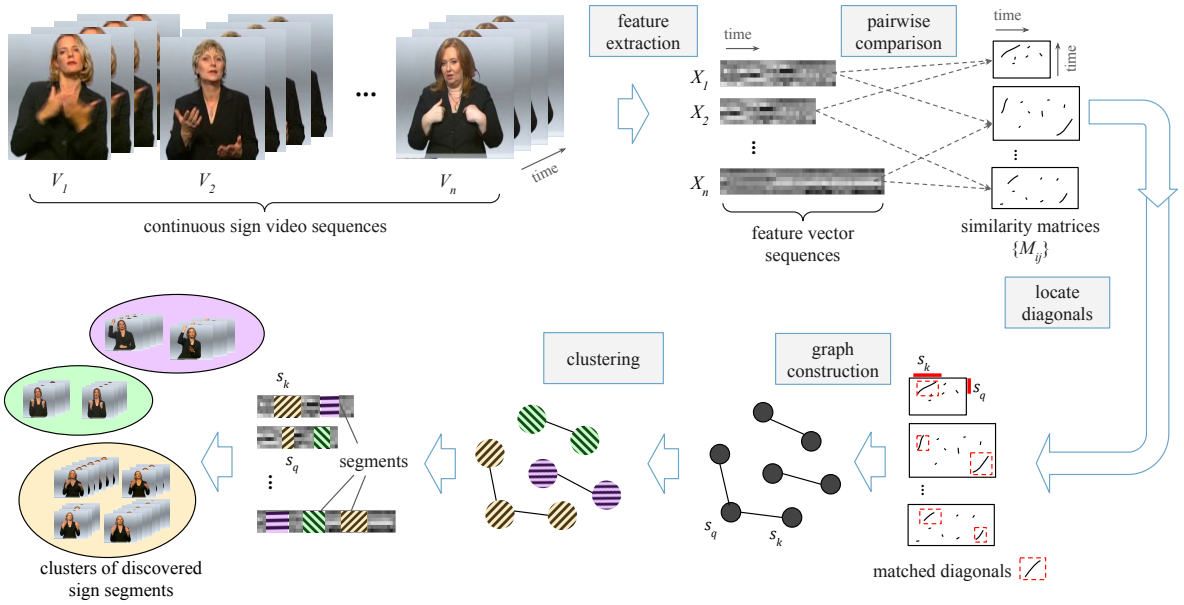


Figure 1: Flow of the unsupervised term discovery algorithm.

curring signs from continuous sign sequences given the information that how many signs are common among these sequences. They report the system performance based on localization of most common signs in 155 sentences from American Sign Language. Even though they do not use labels, their work differs from ours since they use the knowledge of how many sign segments should be discovered from each sequence beforehand. However in UTD, we do not know whether any two sequences share a common sign or not. From the perspective of sub-unit representation, Theodorakis et al. (2014) introduce a sign language phonetic modelling framework in which signs are segmented into dynamic and static sub-units, in an unsupervised fashion. Evaluation of subunit modelling is carried out with regard to ASLR performance on isolated sign datasets. Kelly et al. (2011) and Pfister et al. (2013) use multiple instance learning for extracting isolated signs but utilize text as weak supervision. Our method differs from these in the sense that it does not rely on any supervisory information and analyzes large scale data without knowing whether there are matching segments between pairs of sequences. Our work has the potential for aiding the annotation process for large scale sign dataset when no weak labels (speech, subtitle) are available. It might also be helpful for discovering glosses when little is known about a new sign language.

The rest of the paper is organized as follows. In Section 2 the term discovery algorithm is explained. Experimental setup for sign term discovery is given in Section 3. Implementation details and the results are discussed in Section 4.

2. Term Discovery Algorithm

Our work is based on the algorithm of Jansen and Durme (2011), which is composed of discovery and clustering steps (Figure 1). The discovery step yields pairs of matching segments that are similar to each other. Then an adjacency graph is constructed from the pairs of matching segments and similar segments are clustered together. Details of these steps are explained in the following sections.

2.1. Discovery Step

Starting with a set of feature vectors $\mathcal{X}$, where each vector sequence $X_i \in \mathbb{R}^{d \times T_i}$ is extracted from a continuous signing sequence V_i, the aim is to find pairs of similar sub-sequences by comparing pairs of sequences. For a given pair of feature sequences (X_i, X_j), a pairwise similarity matrix $M_{ij} \in \mathbb{R}^{T_i \times T_j}$ is computed using cosine similarity. If the same word occurs in both sequences, the cosine similarities of the feature vectors corresponding to that time intervals would be high and appear as diagonal lines on the similarity matrix (Figure 1). These regions can be detected by DTW, which is an algorithm that aligns two time series by minimizing alignment costs. By running these steps for all input pairs, we end up with pairs of matching sub-sequences that have low alignment costs.

Both similarity matrix computation and DTW search steps have the time complexity of $O(n^2)$, making it difficult to scale up to large datasets. To combat this limitation, the algorithm of Jansen and Durme (2011) introduces two stages of approximation to perform these steps in $O(n \log n)$ time complexity. These two methods are summarized below.

2.1.1. Approximation of Cosine Similarity Matrix

The first method speeds up the similarity matrix computation by using locality sensitive hashing technique to approximate the pairwise cosine distance computation. At the beginning, feature vectors are normalized such that each dimension has 0 mean and 1 variance across

time. Then applying a random transformation matrix, each vector is projected onto a new 64 dimensional feature space. Projected feature vectors are thresholded at zero to form bit signatures of size 64 (eg. '011...01'). This operation preserves the distance in the original space and enables the approximation of cosine distance by computing the Hamming distance between two bit signatures. Thus if a group of bit signatures are sorted, the signatures that fall nearby would have low Hamming distance between each other. Then the sorted list is linearly swept and for each signature, Hamming distances are computed only between the nearby signatures. This way M is populated sparsely and efficiently without spending time on comparisons that would result in low similarity.

2.1.2. Locating the Diagonal Segments

In the second stage the approximate similarity matrix M is treated as an image. If the same word occurs in both sequences, it would appear as a diagonal line segment on M (Figure 1). These diagonal lines can be located by efficient image processing techniques. First, a diagonal μ-percentile filter of length L is applied which allows the diagonal segments to pass. Then diagonal lines are located with Hough transform which is an algorithm to find lines in an image. Next, segmental DTW search is performed only in the vicinity of the located diagonals instead of exhaustive search over M. The segmental DTW search is terminated when the alignment cost exceeds a threshold C. Using these alignment costs, similarity scores are assigned to matching pairs.

2.2. Clustering Step

As a first step, the matching pairs that have a similarity score less than S_{dtw} are discarded. Using the remaining segments, a graph is formed such that each node corresponds to a discovered segment and the vertices are assigned between the pairs of matching segments. The graph is *de-duplicated* by eliminating overlapping segments. Finally, connected components are found as individual clusters. These clusters are the hypothesized word types and segments are the word tokens.

3. Experimental Setup for Sign Language

The UTD algorithm described above (Jansen and Durme, 2011) is applied to RWTH-PHOENIX-Weather 2014 (Koller et al., 2015) continuous sign dataset by feeding visual features instead of speech features. We implemented some of the metrics that are used in the ZR Challenges (Versteegh et al., 2015; Dunbar et al., 2017) to measure the performance of the UTD algorithm.

3.1. Corpus

A continuous sign dataset which includes gloss annotations with corresponding time boundaries is needed to evaluate this algorithm. In order to satisfy these requirements, we opted to use the RWTH-PHOENIX-Weather 2014 corpus (Koller et al., 2015) which consists of German Sign Language (DGS) interpretation of daily weather forecast on public television, signed by 9 different signers in total. Each video clip is a sign sentence and the glosses for each sentence are annotated manually. However, these manual annotations do not specify the time boundaries of glosses. Follow up work of Koller et al. (2017) uses a Hidden Markov Model (HMM) based forced alignment procedure to find the gloss time boundaries automatically. We used the training set of Multi-Signer setup, since it is the only subset that contains these time boundaries for gloss annotations. We take these automatic annotations as the ground truth labels and use them only for evaluating the performance of the algorithm; the labels are not part of the UTD algorithm.

Working on this corpus leads to several advantages for our task. One advantage is that there are no significant illumination, angle or scale variations. All videos are recorded in the studio with 25fps and signers face directly to a stationary camera. Another advantage is the sign vocabulary being limited to weather related terms only. This results in limited search space for the algorithm and also less variation in signing of a word type. However one drawback is the low resolution (210x260 pixels) of the recordings, which introduces noise to feature extraction process.

Signer ID	Duration (min)	# Sentences	# Discoverable	
			Types	Tokens
1	130	1475	462	15928
5	125	1296	445	13795
4	82	836	345	7642
8	64	704	327	7242
7	60	646	390	7493
3	45	470	260	5227
9	17	165	203	1763
2	6	49	111	576
6	3	30	69	307
Total	533	5671	803	60927

Table 1: Corpus statistics for training set of RWTH-PHOENIX-Weather Multi Signer dataset. A sign type is discoverable if it occurs two or more times.

Corpus statistics for different signer subsets are given in Table 1. We partitioned the Multi-Signer training set further into three subsets, rather arbitrarily. First subset (signer IDs 3,7) constitutes 20% of the training set and it is used as a development set for tuning the parameters of the UTD algorithm. Another subset (signer IDs 2,4,5,6,9) that covers 45% of the training set is used for training the auto-encoder model. The rest (signer IDs 1, 8) are used as the unseen test set for performance evaluation. Note that we used the labels of the development set (IDs 3,7) only for parameter tuning and model validation.

3.2. Features

Convolutional neural nets (CNN) have shown great success in the recent years. Last layers of CNN's capture high

level visual information and their activations can be used as image features. We considered two different feature extraction methods; activations of a pre-trained hand shape classifier CNN and a pose estimator. Moreover, an autoencoder is trained with pose features and the embeddings from the encoder part are used as the third set of features. Our aim is to explore how different types of features can be used for the UTD task, rather than to make a comparison of their performance.

3.2.1. 2D Pose Estimates

OpenPose pose estimator (Cao et al., 2017) is used for obtaining body and hand keypoint coordinates. We used the 8 upper body joint locations out of 25 body joints in addition to 21 keypoints for each hand. Each location is identified with (x, y) coordinates thus we end up with 100 dimensional feature vectors for each video frame. Normalization is done by taking the neck and wrist locations as origins for body and hands respectively and dividing the features by shoulder lengths.

3.2.2. Pre-Trained Hand Shape Classifier

We use the DeepHand convolutional network (Koller et al., 2016) which is a publicly available pre-trained model. Given right hand patches as input, it is able to classify 60 hand shapes based on SignWriting (Sutton, 2000) notation. The final layer is a softmax layer which normalizes the activations to class-conditional posterior probabilities. For our purposes, using the pre-softmax activations as feature vectors is more applicable since distance the UTD algorithm approximates cosine distance between feature vectors. We further applied PCA for dimensionality reduction. We used the wrist coordinates estimated by OpenPose (Cao et al., 2017) to crop right hand patches.

3.2.3. Autoencoder

Autoencoders are encoder-decoder type neural networks which are trained to predict its input. The challenge comes from the existence of a bottleneck layer in the middle, whose dimension is lower than the input dimension. Therefore the network is forced to learn a more compact representation of its inputs, such that from this representation it should be able to reconstruct the original input. They can be formed in varying depths by stacking layers. Here, we make use of this architecture for the purpose of learning better feature representations as well as the additional benefit of achieving non-linear dimensionality reduction. Specifically, we used this network to augment the pose features, aiming for a feature representation that is more appropriate for computing cosine similarity. An autoencoder is trained using the 100 dimensional OpenPose features from the training set. Then using this trained network, the bottleneck features (encoder outputs) are extracted for the development and test sets.

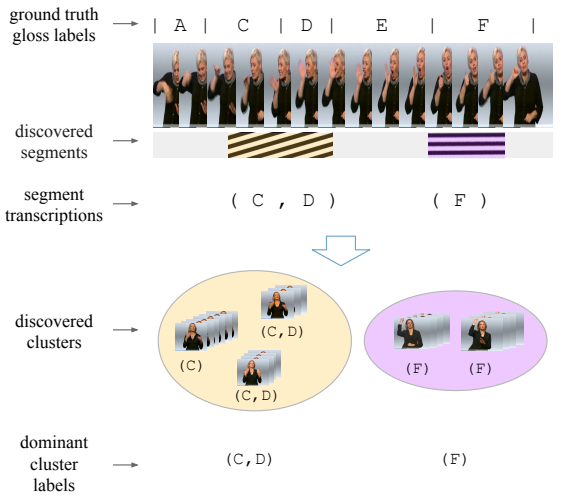


Figure 2: During the evaluation stage, the discovered segments are assigned transcriptions based on the overlapping gloss annotations. Then the metrics are calculated using the transcriptions for each segment. (Gloss labels and images are representative.)

3.3. Evaluation Metrics

Evaluation metrics that we use are inspired by the ZR Challenge (Versteegh et al., 2015; Ludusan et al., 2014). The evaluation for spoken language UTD is based on phoneme level transcriptions however the gloss labels provided by the dataset are not equivalent to phoneme level units. Therefore we modified the evaluation scheme to be compatible with the gloss level annotations.

For the dataset that we use, the labels with time boundaries are aligned using an HMM based model. Koller et al. (2017) used three state HMM's, which model a sign gloss with three sub-units. Therefore the labels they provided indicate the sub-unit indices but we ignore the sub-unit indices and consider only the whole gloss (e.g. $\texttt{WOLKE}_{start}$, $\texttt{WOLKE}_{mid}$, $\texttt{WOLKE}_{end}$ are treated as WOLKE). The annotations also contain *garbage labels* (denoted as 'si' for silence), which might correspond to movement epenthesis and therefore we do not consider it as a target term in evaluation. A discovered segment is mapped to a sequence of ground truth labels if the segment covers at least 50% of that label (see Figure 2). The metrics explained below are calculated using this transcription scheme.

Coverage: The total duration of non-overlapping discovered segments to the duration of all discoverable target segments in the dataset. A target segment contains a gloss that is repeated more than once in the dataset.

Cluster purity: This is a metric that is commonly used to evaluate clustering algorithms. Each cluster is mapped to the most common sequence of ground truth labels. Then purity is the ratio of the segment transcriptions that agree with their dominant cluster label to all discovered

Figure 3: Example clips of three segments that are clustered together.

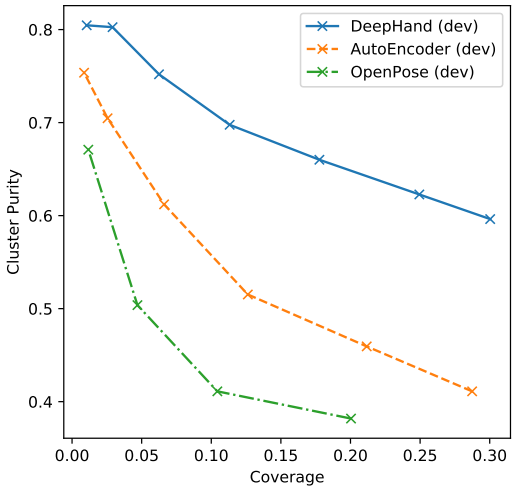


Figure 4: Coverage vs purity curves of sign term discovery using different features in the development set.

segments. For example, if most of the segments in a cluster are mapped to `(C,D)` label and another segment from the same cluster is mapped to `(C)`, it will be penalized due to not having the dominant cluster label. So the purity for the two clusters shown in Figure 2 would be 4/5, since 4 of the segment transcriptions out of 5 agree with the dominant cluster label. More than one cluster may be mapped to the same label, allowing many-to-one mappings.

Even though cluster purity is not included in ZR challenges, here we implement this metric because it is simpler and gives a more intuitive understanding about the quality of clusters. The following metrics are the ones that we implemented for the purpose of enabling comparison with the unsupervised spoken term discovery results. They are computed in terms of precision, recall and their geometric mean (F-scores). Detailed explanations regarding the calculations can be found in (Ludusan et al., 2014).

Matching quality: A set of metrics that measures how well the pairs of segments within a cluster match in terms of substring completion of their transcriptions. For example if transcription for a pair of segments is `(A,B)` and `(A,B,C)`, the substring matches `(A)`, `(B)`, `(A,B)` will be counted as positive for matching precision. Recall is computed over all possbile substring matches that are discoverable.

Grouping quality: This set of metrics measure the inherent quality of the clustering algorithm. It is a similar metric to cluster purity but here it is computed over pairs of segments that belong to discovered clusters instead of single segments. If the pairs of segments that are in the same cluster have the same transcription the precision is high. If inter-cluster pairs have different transcriptions, then the recall is high.

4. Experiments

Hand shape and pose features are extracted for all subsets. Then using the pose features belonging to the train set, an autoencoder model is trained which is then used to obtain the bottleneck features. For each of these three types of features, optimum parameters are found in the development set by grid search over the parameter values. These optimum values are fixed and the system is evaluated on the unseen test set.

When deciding on the autoencoder architecture, we experimented with various bottleneck dimensions (32, 64, 128), hidden unit sizes (64, 128, 256) and also depths (2, 3, 4). Based on the experiments in the development set, the best model has the bottleneck dimension of 64, 128 hidden units and 3 layer depth for both encoder and decoder parts. We trained this model with Adam optimizer (learning rate=0.02).

4.1. Parameter Tuning

The performance of the UTD algorithm (Jansen and Durme, 2011) depends highly on the parameters and the types of features. The default values used in the spoken UTD does not work for sign language because the sampling frequency, average term lengths and the feature properties are different for these domains. Hence, for each type of features (DeepHand, OpenPose, AE), we run grid search over the parameters only in the development set and pick the best combination of parameter values using the evaluation setup. The term discovery results on the development set are shown in Figure 4. These curves are obtained by sweeping the score threshold S_{dtw}, according to which the pairwise matches are eliminated before the graph clustering step. The curves illustrate the trade of between coverage and purity. The optimum parameters would yield both high coverage and high purity, and

Set	Feature Type	Discovered		Coverage (%)	Matching (%)			Grouping (%)		
		Clusters	Segments		Precision	Recall	F-score	Precision	Recall	F-score
Test	DH	858	2208	12.8	32.5	3.9	6.9	45.0	56.5	50.1
	AE	606	1506	10.4	8.0	1.3	2.3	12.9	23.0	16.5
	OP	181	384	5.4	0.0	0.0	0.0	1.6	8.8	2.7
Dev	DH	179	663	10.5	24.6	4.3	7.4	51.4	69.3	59.0
	AE	210	922	13.7	3.4	3.8	3.6	33.9	53.1	41.4
	OP	94	527	9.8	1.1	3.2	1.6	32.8	61.3	42.8
ZR'15 baseline (Eng)				16.3	39.4	1.6	3.1	21.4	84.6	33.3

Table 2: Sign discovery results obtained using different features (DH: DeepHand, AE: autoencoder, OP: OpenPose) in the development and test set. The baseline results of Zero Resource spoken term discovery challenge are given in the bottom row for comparison.

comparison of different setups can be deduced visually from these curves.

Even though the optimum values of parameters for each feature type vary, they are not much different from each other and here we share a combination that generally yields good results in this setup. The optimum values for the parameters that are described in Section 2.1.2 are as follows: for the percentile filter $L = 11$, $\mu = 0.60$ and for cost threshold $C = 4$.

4.2. Test on Unseen Signers

Using the optimum parameters for each feature type, we run experiments on the test set. We selected the S_{dtw} threshold such that coverage is around 10% and evaluated the discovery results as shown in Table 2. The hand shape features yield the best results for each setup. This might be because the dataset we use is one of the three datasets that DeepHand is trained over. It is trained with hand shape class labels, not the gloss information but nevertheless, having seen these images before might explain the robust performance on this dataset. Using the autoencoder resulted in slight improvements over the pose features both in the development and test sets. It might be the case that the bottleneck features provide a better representation when using cosine similarity for pairwise comparisons. Poor performance of the pose features is not because they are inferior to hand shapes. It can be due to low resolution of the images but more probably, it might not be applicable to compare two set of joint coordinates with cosine similarity. A feature transformation that is more relevant to similarity comparisons should be applied to pose features. We show that even a simple neural net can improve the pose features and more complex models would probably boost the performance. Pose features can be processed by graph convolutional encoders to better capture the connectivity relationships and temporal dependencies between joints. As an exploration of possibilities, here we aim to demonstrate that different types of features can be tailored to achieve better results for this task.

Using DeepHand on the test set, the glosses that are found most accurately are given in Figure 5. The occurrences represent the number of times the gloss type is clustered

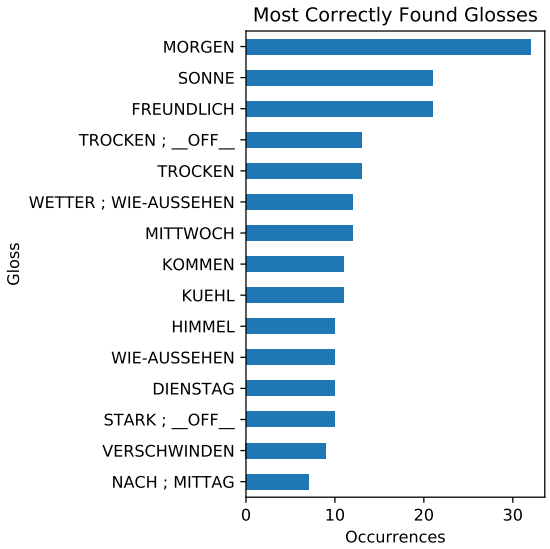


Figure 5: The glosses that the algorithm is able to cluster with %100 purity.

with 100% purity. Here some of the discovered types consist of two or more consecutive glosses and it shows that the algorithm can discover sequences of glosses (n-gram) if they occur frequently. It is observed that it can discover up to 4-gram. In Figure 6, some of the most confused gloss types are shown. The numbers between pairs of glosses are the number of co-occurrences in a cluster that has less than 50% purity. These words that are confused with each other have similar semantics and almost identical signings except for the mouthing. This suggests the importance of incorporating non-manual features to the feature extraction.

4.3. Discussion and Future Work

The proposed approach can aid sign language community in numerous ways. First of all, it can be very useful in cases where there are large amounts of sign videos to be annotated but not enough available resources. The algorithm proposes segments and clusters them so that each cluster corresponds to a hypothesized gloss. An ed-

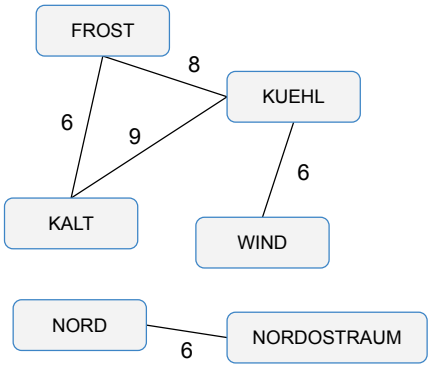


Figure 6: This graph shows some of the most confused gloss pairs. The numbers indicate how many times the pair is grouped together in a cluster that has less than 50% purity.

ucated annotator can easily purify the discovered clusters by eliminating the false segments and then saving the segments from pure clusters as annotations. By doing so, a significant amount of data can be annotated in short time. Given the pre-computed features, 1 hour video is processed in about 15 minutes by a 16-core CPU and an annotator can review the discovered clusters quickly with a proper software. However, the quality of the segmentation boundaries is not assessed in this work. In a future study, a psycho-linguistic experiment can be carried out, where the subjects are shown signs that are segmented by humans versus the UTD algorithm and they are asked to decide on which segmentation seems more natural. This would validate the potential use case of the UTD method as an automatic segmentation tool. Another benefit may arise when we want to train an ASLR system on a sign language that does not have enough resources. The clusters found by the UTD algorithm can provide weak supervision for training models, such as correspondence autoencoders proposed by Kamper et al. (2015). Finally researchers can use this as a tool to build lexicon for a new sign language.

Most of the clusters contain 2 or 3 segments. This is caused by the way the adjacency graph is clustered. As the clustering algorithm, connected components method simply thresholds the edge weights and groups the nodes that remain connected. However, using a more sophisticated algorithm (eg. modularity based), some of the similar clusters can be further joined, hence grouping recall can be increased. Analysis of such algorithms for UTD is done by Lyzinski et al. (2015) and the comparison of these algorithms on unsupervised sign discovery can be a subject of future study.

One of the drawbacks of this work is having used only one corpus for development and testing. Although the testing is done on unseen signers, the language is the same and the recording conditions are almost identical. A future study may include another sign corpus with a different sign language for testing. This would enforce the system to be language independent and would require better feature representations that can generalize well.

5. Conclusion

In this paper, we introduce a new task for sign language processing; unsupervised sign term discovery. The aim is to discover gloss types by clustering segments from a continuous sign dataset using only the video signal. We show that a highly acclaimed spoken term discovery algorithm can be run on continuous sign language videos by using visual features. The results show that, using appropriate features, the algorithm can achieve similar performance compared to its application in the spoken language domain. We believe that the studies targeting this task will lead to better annotation and ASLR systems in the future.

6. Acknowledgements

This work is supported in part by the Scientific and Technological Research Council of Turkey (TUBITAK) under Project 117E059.

7. Bibliographical References

Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. (2017). Realtime multi-person 2D pose estimation using part affinity fields. In *Proc CVPR*, pages 1302–1310.

Dunbar, E., Cao, X., Benjumea, J., Karadayi, J., Bernard, M., Besacier, L., Anguera, X., and Dupoux, E. (2017). The zero resource speech challenge 2017. In *Proc. ASRU*, pages 323–330, Dec.

Jansen, A. and Durme, B. V. (2011). Efficient spoken term discovery using randomized algorithms. In *Proc. ASRU*.

Kamper, H., Elsner, M., Jansen, A., and Goldwater, S. (2015). Unsupervised neural network based feature extraction using weak top-down constraints. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5818–5822, April.

Kelly, D., Mc Donald, J., and Markham, C. (2011). Weakly supervised training of a sign language recognition system using multiple instance learning density matrices. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 41(2):526–541, April.

Koller, O., Forster, J., and Ney, H. (2015). Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125, December.

Koller, O., Ney, H., and Bowden, R. (2016). Deep hand: How to train a CNN on 1 million hand images when your data is continuous and weakly labelled. In *Proc. CVPR*, pages 3793–3802, June.

Koller, O., Zargaran, S., and Ney, H. (2017). Re-Sign: Re-Aligned End-to-End Sequence Modelling with Deep Recurrent CNN-HMMs. In *Proc. CVPR*, pages 3416–3424, July.

Ludusan, B., Versteegh, M., Jansen, A., Gravier, G., Cao, X.-N., Johnson, M., and Dupoux, E. (2014). Bridging the gap between speech technology and natural language processing: an evaluation toolbox for term discovery systems. In *Proc. Language Resources and Evaluation Conference*, May.

Lyzinski, V., Sell, G., and Jansen, A. (2015). An evaluation of graph clustering methods for unsupervised term discovery. In *Proc. Interspeech*.

Nayak, S., Duncan, K., Sarkar, S., and Loeding, B. L. (2012). Finding recurrent patterns from continuous sign language sentences for automated extraction of signs. *Journal of Machine Learning Research*, 13:2589–2615.

Park, A. S. and Glass, J. R. (2008). Unsupervised pattern discovery in speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):186–197, Jan.

Pfister, T., Charles, J., and Zisserman, A. (2013). Large-scale learning of sign language by watching TV (using co-occurrences). *Proceedings of the British Machine Vision Conference*, pages 1–11.

Sutton, V. (2000). Sign writing. *Deaf Action Committee (DAC) for Sign Writing*.

Theodorakis, S., Pitsikalis, V., and Maragos, P. (2014). Dynamic-static unsupervised sequentiality, statistical subunits and lexicon for sign language recognition. *Image Vision Comput.*, 32:533–549.

Versteegh, M., Thiollière, R., Schatz, T., Cao Kam, X.-N., Anguera, X., Jansen, A., and Dupoux, E. (2015). The zero resource speech challenge 2015. In *Proc. Interspeech*, pages 3169–3173.

Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 197–202
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

The Corpus of Finnish Sign Language

Juhana Salonen, Antti Kronqvist, Tommi Jantunen
University of Jyväskylä, Sign Language Centre, P.O. Box 35, FI-40014 University of Jyväskylä, Finland
{juhana.k.salonen, antti.r.kronqvist, tommi.j.jantunen}@jyu.fi

Abstract

This paper presents the Corpus of Finnish Sign Language (Corpus FinSL), a structured and annotated collection of Finnish Sign Language (FinSL) videos published in May 2019 in FIN-CLARIN's Language Bank of Finland. The corpus is divided into two subcorpora, one of which comprises elicited narratives and the other conversations. All of the FinSL material has been annotated using ELAN and the lexical database Finnish Signbank. Basic annotation includes ID-glosses and translations into Finnish. The anonymized metadata of Corpus FinSL has been organized in accordance with the IMDI standard. Altogether, Corpus FinSL contains nearly 15 hours of video material from 21 FinSL users. Corpus FinSL has already been exploited in FinSL research and teaching, and it is predicted that in the future it will have a significant positive impact on these fields as well as on the status of the sign language community in Finland.

Keywords: Corpus of Finnish Sign Language, Language Bank of Finland, Finnish Signbank, annotation, metadata, research, teaching

1. Introduction

This paper presents the Corpus of Finnish Sign Language (Corpus FinSL). The corpus was published in FIN-CLARIN's Language Bank of Finland in May 2019 as a result of the four-year (2014–2018) CFINSL (Corpus of Finland´s Sign Languages) project led by the University of Jyväskylä, Finland (see Salonen et al., 2016).[1] The aim of the CFINSL project was to collect, process and make openly available narratives and conversations in FinSL and Finland-Swedish Sign Language (FinSSL), which are the two official sign languages in Finland. During the project, video material was recorded from 91 FinSL users and 12 FinSSL users of different ages from all over Finland.

The published material includes only FinSL data and comprises nearly 15 hours of signing recorded in 2014 from 21 signers. The material has been annotated in ELAN[2] (Max Planck Institute in Nijmegen; Crasborn & Sloetjes, 2008) for signs and Finnish translations. In addition to this, the published material includes anonymized metadata structured according to the IMDI standard (see Section 3). All of the published material is available for academic use under two licenses: a part of the data (elicited narratives) is licensed with Creative Commons 4.0 BY-NC-SA and a part (conversations) with FIN-CLARIN RES (see Table 1 and Section 2). A summary of the content of the published corpus material is shown in Table 1.

All of the material	14 hours and 22 minutes
Elicitated narratives (CC)	5 hours and 4 minutes
Conversations (RES)	9 hours and 18 minutes
Video files	343 mp4-files
Annotation files	142 files (eaf +pfsx)
Number of informants	21 informants

Table 1: Statistics of Corpus FinSL in the Language Bank of Finland.

2. The Subcorpora and Their Motivation

Corpus FinSL in the Language Bank of Finland has been divided into two subcorpora on the basis of licenses and the content of the video material: Elicited narratives and Conversations. The Elicited narratives (Corpus FinSL-elicit) are publicly available for researchers, educators and the general public under the newest non-commercial Creative Commons license (CC 4.0 BY-NC-SA).[3] These narratives consist of signed retellings of cartoon strips, videos and picture books. Access to the dataset of Conversations (Corpus FinSL-conv) is restricted to academic use only; it requires a research plan as well as personal access rights, in accordance with the RES license of the Language Bank of Finland.[4] The material comprises conversations about different topics such as work and hobbies as well as deaf culture. The whole data consists of seven different tasks, which are presented in Table 2. The data of Corpus FinSL-elicit contains tasks 3, 4 and 5 (marked with an asterisk) while the data of Corpus FinSL-conv includes tasks 1, 2, 6 and 7. The structure of Corpus FinSL in the Language Bank of Finland is shown in Figure 1.

Task types:
(1) presenting oneself
(2) telling about one's hobby/work
*(3) signing cartoon strips (Ferd'nand)
*(4) signing a video story (Mr. Bean and Laurel & Hardy videos)
*(5) signing from a picture book (*The Snowman* and *Frog, Where Are You* picture books)
(6) discussing an event related to Deaf culture
(7) free discussion

Table 2: The task types of Corpus FinSL. Tasks marked with an asterisk are freely available whereas tasks without an asterisk are restricted to academic use only.

[1] http://r.jyu.fi/tTc
[2] https://tla.mpi.nl/tools/tla-tools/elan/
[3] https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode
[4] https://www.kielipankki.fi/support/clarin-eula/

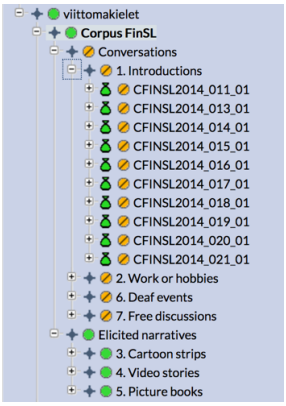


Figure 1: Screenshot showing the structure of Corpus FinSL in the LAT platform of the Language Bank of Finland.

The division of Corpus FinSL material into two subcorpora is primarily due to compliance with the EU Data Protection Regulation (GDPR), which came into force in May 2018. Whereas the subcorpora of Elicited narratives contains only thematically restricted monologues, in which the informants replicate the stories of different picture and video materials, the subcorpora of Conversations includes more thematically free dialogues, in which the informants may also indirectly disclose personal information from third parties. Permission is never granted for the free disclosure of this personal information, so access to the conversation data has been restricted. The decision was made in collaboration with the legal department of the University of Jyväskylä and FIN-CLARIN.

In general, the informants' consent was collected in two phases. First, after a filmed session (in 2014), the informants were asked to choose either the Yes or No option for each of the following five statements:

1. Video material can be used for research purposes, but publishing video clips or still images is prohibited
2. Video material can be presented in public events (e.g. academic presentations and teaching)
3. Still images can be taken from the video material for publications (electronic or paper)
4. All the video material can be published electronically e.g. in the Internet
5. The name of the participant can be mentioned in publications.

When the publication of the material became imminent (in 2018), after the GDPR had come into force, it was necessary to update the consents originally collected with additional information. While the original consent treated the online publication of the material as one general entity, the additional consent form that we devised required the informants' explicit permission for the long-term preservation of the material in the Language Bank of Finland and the free publication of each communication assignment (seven task types of varying nature and privacy) on the same platform. Pursuant to the GDPR´s guidelines, the informants were also asked to explicitly consent to the inclusion of personal data belonging to certain specific categories. In particular, this concerned the information they had given us with respect to their hearing status, which falls into the category of information concerning special groups mentioned in the GDPR. This type of data was included not only in the background forms collected from the informants but also, for example, arises in interactive task 1, where the informants naturally identify themselves as deaf when presenting themselves in a way specific to the sign language culture.

In the additional consent form, the informants were asked to choose, as in the original consent form, either the Yes or No option for questions 1, 2 and 4. The third question, on the other hand, included a list of all seven task types, of which the informants then selected those which they allowed to be published. Before completing the consent form, the informants received all of the necessary information in both Finnish and FinSL.

1. In addition to my prior consent (dated X) in the attachment, I consent to the licensing of the recording that was the subject of this consent, under a *Creative Commons-ByAttribution-NonCommercial-ShareAlike 4.0 International* (BY NC SA 4.0) license, and agree that the University of Jyväskylä is the copyright holder of this recording
2. In addition, I consent to the transfer for long-term preservation of the licensed recording that was the subject of this consent to FIN-CLARIN's Language Bank of Finland, where it may be used for research purposes on the basis of appropriate research plans
3. I also consent to the publication of the licensed recording that was the subject of this consent in the Language Bank for the tasks that I have ticked below, meaning that anyone can see and use them
4. I consent to the inclusion in the recording of information that falls into specific categories of personal data, such as the degree of my hearing

In practice, the additional consent form also required informants to consent to the licensing of their entire material under the Creative Commons 4.0 BY NC SA license, which ultimately prohibits commercial use. However, as explained above, only the Elicited narratives subcorpora is covered by this license. Free publication of the Conversation subcorpora is possible in the future if the annotations and video material comprising it are anonymized for any third party related information.

3. Annotations and Metadata

The published Corpus FinSL contains over 107,000 glossed sign tokens. The annotation process (Sections 3.1 and 3.2) began with the identification of sign units, the distinction of their meanings and forms (see Salonen et al., 2016) as well as provisional information about the grammatical behavior of the sign (e.g. negation), and the

translation of chunks of signed utterances into Finnish. The metadata (3.3) describes the generic nature of the dataset, the individuals present in the collection, the content of the videos, and the formats of the video and annotation files.

3.1 Sign Level Annotation

Unity and consistency are the prerequisites for building a functioning corpus. This means that common principles and annotation guidelines must be agreed between all annotators. Annotation conventions[5] have been developed in several sign language corpus projects (e.g. Johnston, 2016 Australia; Schembri et al., 2013 the United Kingdom; Crasborn et al., 2015 the Netherlands; Wallin & Mesch, 2018 Sweden). In the CFINSL project, we began to develop the conventions during the basic annotation (see Keränen et al., 2016). The first version of the annotation conventions was released in spring 2018 and included the guidelines for sign level annotation in the CFINSL project (see Salonen et al., 2018 as well as Tables 3 and 4). There are guidelines in the conventions for the recording of different parts of the lexicon, such as lexical signs of varying degrees (see Jantunen, 2018). Guidelines related to creating sentence level translations were added to the second version of the conventions, published in February 2019 (Salonen et al., 2019).

Category	Example
Lexical signs	a common/distinct ID-gloss:
○ Phonetic variants (1-2 different parameters)	FATHER(Ax), FATHER(G) => FATHER
○ Lexical variants (2-3 different parameters)	with e.g. a handshape code: DO(BB) vs. DO(SS)
○ Polysemic signs	BALL, WORLD => BALL
○ Homonym signs	BEACH, BORDER => BORDER
Numeral signs (_num)	SIX-YEAR_num, ONE-WEEK-EARLIER_num
Pointings (OS:)	OS:, OS:me
Depicting signs (_kv + a subclass code)	_kvkk (a whole entity classifier)
Gestural signs (_ele)	PALM-UP_ele
Fingerspellings (_sa)	t-o-m-i_sa

Table 3: Examples of the conventions of sign level annotation.

The finalized sign-level annotation of Corpus FinSL material exploits ID-glosses, which means that signs of the same form (homonymous, polysemic, and phonetically variable) are assembled under the same tags (for the annotation process more generally, see Salonen et al. 2016). In practice, an ID-gloss refers to a label selected to represent sign tokens that have a similar form but whose meanings may vary in a corpus (Johnston, 2008, 2010). For example, in FinSL there is a manually articulated sign, the same form of which can mean 'everyday', 'jeans', 'rural', 'fresh' or 'orange', depending on the context in which it appears. Instead of tagging the tokens with a context-specific meaning gloss, the sign is glossed with a single label EVERYDAY to represent all tokens of the same form. Thus the ID-gloss does not indicate a meaning translation of the sign, but rather functions as an identifier agreed upon by the annotators. ID-glossing allows one to search the data more efficiently than glossing according to a contextual meaning. The basic annotation of Corpus FinSL has been focused on identifying sign tokens as much as possible in a systematic and fast manner from the perspective of the annotators without actual research guidelines.

Type of grammaticality	Code
Negation	@neg
Repetition+plural	@toisto
Compound sign	@y
List buoy	@poijul
Lexicalized fingerspellings	@sv
Foreign borrowings	@lv

Table 4: Examples of the conventions for the grammatical behavior of the sign.

We implemented ID-glossing on two interconnected platforms: glosses that are temporarily connected to video material in ELAN are also connected to a database of Finnish Signbank[6] via a network connection. Finnish Signbank is the lexicon database built for FinSL and FinSSL; its basic function is to serve as a tool for annotating sign language texts. In addition to the gloss, Finnish Signbank contains the citation form of the sign on video(s), the sign´s Finnish or Swedish equivalents, and any further information on the sign, if necessary. The database can be updated as annotation work progresses. Figure 2 contains four videos of the same sign form but with different meanings. The difference in meaning can be detected, for example, with the help of a mouthing.

The ELAN program includes a feature that allows the program to access external controlled vocabulary (ECV) maintained by an external web server when annotating. In the CFINSL project, a controlled vocabulary (CV) was placed in Finnish Signbank by developing the Signbank platform for this purpose. The CV allows the annotator to

[5] http://r.jyu.fi/ygR

[6] https://signbank.csc.fi/

check whether the gloss is already in the database when labeling the annotation cell. If the gloss already exists, the annotator can select it from the list, thus avoiding spelling mistakes, which are easily generated in non-automated transcription. If the gloss has not yet been created for that sign, one can add it to the database and create a new gloss record with videos, translations, and more. Figure 3 illustrates how the ELAN program makes it possible, when creating annotation cell content, to search for a suitable gloss from Finnish Signbank with the help of either the ID-gloss (left-hand column of the box) or its translation equivalents (right-hand column of the box). If necessary, it is also possible to check the video(s) on the gloss page of Signbank. (Salonen et al., 2018.) Manually executed gloss and translation changes in Signbank are automatically updated on all linked annotation cells with a continuous ECV connection.

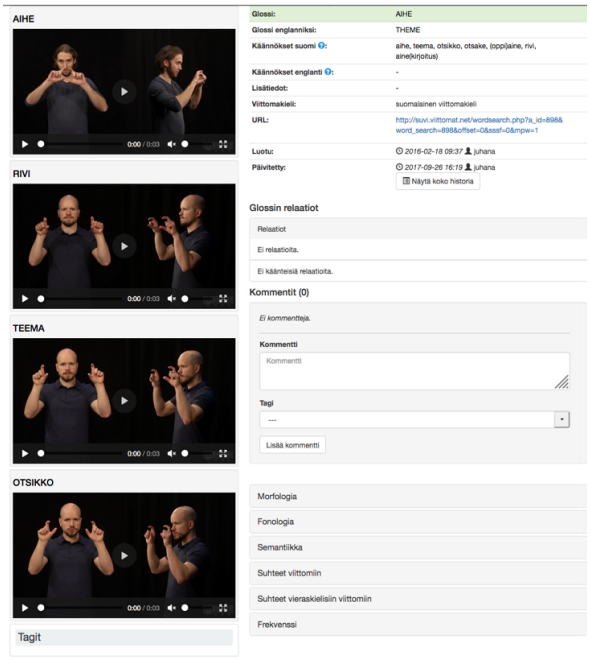


Figure 2: A view of a gloss page in Finnish Signbank.

Finnish Signbank has been developed by the CFINSL project, which has been cooperating with the corpus and dictionary work of the Finnish Association of the Deaf. The database is based on Auslan Signbank,[7] originally developed for Australian Sign Language, and its subsequent application, the Dutch Signbank database. Source codes from all Signbank databases are available on the Github version control site,[8] where the database structure and features of Finnish Signbank can also be found, documented in a user-friendly way (named FinSL-signbank wiki). Signbank development started some ten years ago with international collaboration between research teams in Australia, the Netherlands, Finland and the United Kingdom (Cassidy et al., 2018). Finnish Signbank comprises two different and independent lexicons. The lexicon of Corpus FinSL contains all of the fixed signs (i.e. lexical signs), as well as a small group of depicting, gestural (emblems) and pointing signs (for further analysis) that have been found in the corpus material. In addition, there is the lexicon of the Kipo corpus of the Finnish Association of the Deaf, which is based on annotated material (approximately 2.5 hours) from the Language Policy Programme for the National Sign Languages in Finland (Kuurojen Liitto ry, 2015).

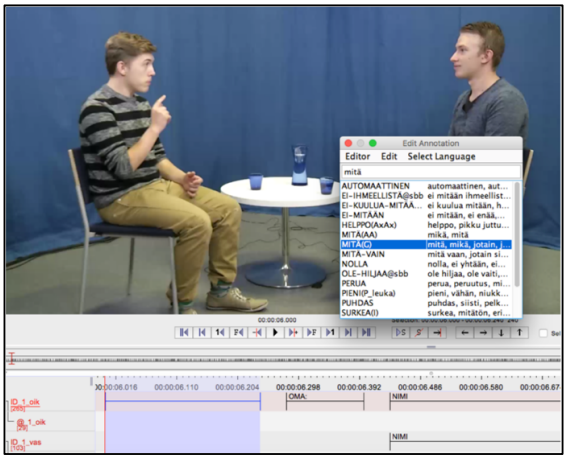


Figure 3: A view of annotation in ELAN using a controlled vocabulary hosted by Signbank.

3.2 Sentence Level Annotation

The video material has also been annotated on the sentence level. In practice, this means that the signing has been translated into Finnish. At the beginning of the translation, meaningful sentences were separated from the referenced text stream by the translators' intuition, without further distinction of sentences, which is the task of the actual study after basic annotation. The translation is in a form that takes into account the way in which the source language is expressed, both manually and non-manually (with the head, body and face).

In addition, to the translations have been added parts in parentheses which a fluent Finnish text requires, but which are not made visible in the preceding discourse context or which may not be required at all in sign language text (e.g. the subject of the sentence, a copula, some conjunctions; see Example 1). The translation guidelines are described in more detail in the annotation conventions (see Salonen et al., 2019).

(1) LOOK-AT OUTSIDE RAIN SNOW RAIN
(He/she) notices, (that) it is snowing outside.

The translation provides a more complete view of the signed texts, as ID-glossing focuses solely on manual articulation. The translation can also be used to check what

[7] http://www.auslan.org.au/

[8] https://github.com/Signbank

meaning the ID-gloss is referring to. The translators have also made Finnish translations of signs for gloss pages in the Finnish Signbank database, according to the meanings of the signs marked with ID-glosses in connection with the translation.

3.3 Metadata

The anonymized metadata of Corpus FinSL transferred to the Language Bank of Finland are described in accordance with the IMDI (*ISLE Meta Data Initiative*) standard. IMDI is a description standard developed at the Max Planck Institute in Nijmegen for consistency in the description of multimedia and multimodal language materials.[9] In accordance with IMDI standards, the CFINSL project produced a general description of the material (*Corpus FinSL*), the underlying project (*CFINSL Project*), and the contents of both of the subcorpora (*Elicited narratives; Conversations*) as well as their communication tasks (1-7). In addition, a session-specific description was made of the individual communication tasks of each pair (*Session*), which included information about the participants (*Actors*); the quality of the communication situation, its interactivity and its collection method (*Content*); video materials (*MediaFiles*); and annotations (*WrittenResources*).

Background information about the informants was collected very extensively during the CFINSL project collection. However, the IMDI description built into the Language Bank eventually selected only an anonymous code that identifies the individual, age group, gender, area of residence, and handedness (left/right).

4. Exploitation of Corpus FinSL in Research and Teaching

The material of the CFINSL project in general and of Corpus FinSL in particular has already been exploited in several research projects focusing on FinSL. First of all, the corpus provided new insights into the study of FinSL word order (e.g. Jantunen 2017), lexicon (Takkinen et al., 2018) and nonmanuality (Puupponen 2018). Currently, the corpus is the main source of data for a larger project which is investigating the role of gesturality in language by focusing on the use and variation of constructed action in FinSL. In addition, the material has been used in comparative studies of sign languages (Jantunen et al., 2016; Puupponen et al., 2016) and in Master's theses completed on the subject of Finnish Sign Language (e.g. Syrjälä, 2018; Puhto, 2018). In general, the existence of the corpus has already had a significant impact on the research tradition of FinSL by requiring that individual studies should be more closely connected to the material that has been collected.

Another, but globally not so recognized, area where Corpus FinSL has made a contribution in the field of FinSL is teaching. Corpus FinSL was taken into account in the planning of the new curriculum for FinSL in Jyväskylä University (2020-2023): the corpus material is included in the course descriptions and targeted as learning outcomes of different courses. This obliges teachers to apply the corpus in teaching.

In addition, in the fall of 2019, in-service training[10] for sign language teachers funded by the Ministry of Education and Culture was started at the Sign Language Centre, University of Jyväskylä. The aim of the in-service training is to keep sign language teachers informed about new research. In-service training includes three different courses: FinSL grammar, Deaf Studies, and (sign) language acquisition and language assessment. We have made new learning materials, especially in the FinSL grammar course, which is based on Corpus FinSL data. At the moment, most sign language teaching materials are not aligned with the newest research findings, especially within corpus-based research.

The learning materials we have created are given to participants on the in-service training course so that they can exploit the materials in their own teaching. We plan in the future to publish the learning materials on a website where they will be freely available to everyone for teaching purposes.

From the experience we have gathered so far of using the corpus in teaching we have seen that the corpus can be used in teaching in at least three different ways.

1. Examining the corpus data; finding and discussing relevant language-related sign units and language structures.
2. Annotation of corpus data by students. This helps to internalize language usage and variation in language units.
3. Searching for corpus data to illustrate a theoretical perspective (e.g. sign types, word order in FinSL); additional annotation may be required.

All this requires appropriate (research) literature; corpus data is used and discussed in relation to the (research) literature. In addition, corpus data can be used to teach deaf culture or sign language communication. The data contains e.g. conversations about events related to deaf culture, and offers interesting topics for discussion. The corpus includes interaction between two persons and therefore also offers an opportunity to analyze various interaction practices.

5. Discussion and Conclusion

In this paper we have described the content of Corpus FinSL as it is published in the Language Bank of Finland. We have presented the annotation guidelines, metadata and exploitation of the corpus material in research and teaching. The extensive electronic and computer-readable material offers new opportunities for quantitative and qualitative research on the sign languages used in Finland; it already does this for FinSL, and later it will do the same for FinSSL. The wide-ranging, multi-person material can be used to examine, for example, the signing of native signers of different ages from around Finland, as well as

[9] https://tla.mpi.nl/imdi-metadata/

[10] http://r.jyu.fi/CyD

differences in signing and language structure between different types of text genres. The extensive, partly publicly available data also allows for a completely new way of comparing sign languages. This is supported in particular by the use of similar data collection methods in corpus projects in different sign languages.

Corpus FinSL will have a significant impact on Finland´s sign language community and the social status of sign languages. For sign language people, it provides an opportunity to develop language awareness of their mother tongue, which many have not been taught in basic education. For those using sign language as a foreign language – such as sign language interpreters – the corpus provides educational material on, for example, recognizing the socio-linguistic differences between language users. The corpus material will also continue to be used in an ongoing project at the University of Jyväskylä to develop in-service training for sign language teachers in Finland. Finally, in addition to its teaching and training applications, the corpus may in the future serve as a tool for developing language management and language planning.

6. Acknowledgements

The authors wish to thank Eleanor Underwood for checking the English of the paper and all people who have participated in building Corpus FinSL, especially professor emerita Ritva Takkinen who took the necessary first step.

7. References

Cassidy, S., Crasborn, O., Nieminen, H., Stoop, W., Hulsbosch, M., Even, S., Komen, E. & Johnston, T. (2018). Signbank: Software to Support Web Based Dictionaries of Sign Language. *Proceedings - The 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community*, 2359-2364.

Crasborn, O. & Sloetjes, H. (2008). Enhanced ELAN functionality for sign language corpora. *Proceedings - The 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, 39-43.

Crasborn, O., Bank, R., Zwitserlood, I., Kooij, E., Meijer, A. & Sáfár, A. (2015). Annotation Conventions for The Corpus NGT. Version 3. Radboud University Nijmegen: Centre for Language Studies & Department of Linguistics.

Jantunen, T. (2018). Viittomakielet hybridisysteemeinä: hämärärajaisuus ja epäkonventionaalisuus osana viittomakielten rakennetta. *Puhe ja Kieli* 38(3), 109-126.

Jantunen, T. (2017). Fixed and NOT free: Revisiting the order of the main clausal constituents in Finnish Sign Language from a corpus perspective. *SKY Journal of Linguistics* 30, 137-149.

Jantunen, T.; Mesch, J.; Puupponen, A. & Laaksonen, J. (2016). On the rhythm of head movements in Finnish and Swedish Sign Language sentences. *Proceedings - Speech Prosody 2016*, 850-853.

Johnston, T. (2008). Corpus linguistics and signed languages: No lemmata, no corpus. *Proceedings - The 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, 82-87.

Johnston, T. (2010). From archive to corpus: Transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics* 15 (1), 106-131.

Johnston, T. (2016). *Auslan Corpus Annotation Guidelines*. Centre for Language Sciences, Department of Linguistics, Macquarie University (Sydney) and La Trobe University (Melbourne), Australia.

Keränen, J., Syrjälä, H., Salonen, J. & Takkinen, R. (2016). The Usability of the Annotation. *Proceedings - The 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining,* 111-116.

Kuurojen Liitto ry (2015). *Suomen viittomakielten kielipoliittinen ohjelma 2010 -korpus, annotoitu versio*. Kielipankki. http://urn.fi/urn:nbn:fi:lb-2014073031

Puhto, J. (2018). *Päänpudistuksen käyttötavat ja frekvenssit suomalaisessa viittomakielessä.* Suomalaisen viittomakielen pro gradu -tutkielma. Kieli- ja viestintätieteiden laitos, Jyväskylän yliopisto.

Puupponen, A., Jantunen, T. & Mesch, J. (2016). The Alignment of Head Nods with Syntactic Units in Finnish Sign Language and Swedish Sign Language. *Proceedings - Speech Prosody 2016*, 168-172.

Puupponen, A. (2018). The relationship between the movements and positions of the head and the torso in Finnish Sign Language. *Sign Language Studies* 18(2), 175-214.

Salonen, J., Takkinen, R., Puupponen, A., Nieminen, H. & Pippuri, O. (2016). Creating Corpora of Finland´s Sign Languages. *Proceedings - The 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining*, 179-184.

Salonen, J., Wainio, T., Kronqvist, A. & Keränen, J. (2018). *Suomen viittomakielten korpusprojektin (CFINSL) annotointiohjeet.* 1. versio. Kieli- ja viestintätieteiden laitos, Jyväskylän yliopisto. http://r.jyu.fi/ygQ

Salonen, J., Wainio, T., Kronqvist, A. & Keränen, J. (2019). *Suomen viittomakielten korpusprojektin (CFINSL) annotointiohjeet.* 2. versio. Kieli- ja viestintätieteiden laitos, Jyväskylän yliopisto. http://r.jyu.fi/ygR

Schembri, A., Fenlon, J., Rentelis, R., Reynolds, S. & Cormier, K. (2013). *Building the British Sign Language Corpus*. Language Documentation and Conservation 7, 136-154.

Syrjälä, H. (2018). *Hakukysymysviittoman paikka suomalaisessa viittomakielessä.* Suomalaisen viittomakielen pro gradu -tutkielma. Kieli- ja viestintätieteiden laitos, Jyväskylän yliopisto.

Takkinen, R., Keränen, J. & Salonen, J. (2018). Depicting Signs and Different Text Genres: Preliminary Observations in the Corpus of Finnish Sign Language. *Proceedings - The 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community*, 189-194.

Wallin, L. & Mesch, J. (2018). Annoteringskonventioner för teckenspråkstexter. Version 7. Avdelningen för teckenspråk, Institutionen för lingvistik, Stockholms universitet.

Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 203–208
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

Tools for the use of SignWriting as a Language Resource

Antonio F. G. Sevilla[1], Alberto Díaz Esteban[2], José María Lahoz-Bengoechea[3]
[1]Knowledge Engineering Institute,
Facultad de Psicología, Lateral 2, Campus de Somosaguas 28223 Pozuelo de Alarcón, Spain
[2]Department of Software Engineering and Artificial Intelligence,
Facultad de Informática, c/ Profesor José García Santesmases, 9 28040 Madrid, Spain
[3]Department of Spanish Linguistics and Literary Theory,
Facultad de Filología, edificio D, c/ Prof. Aranguren s/n, 28040 Madrid, Spain
Universidad Complutense de Madrid
[1]afgs@ucm.es, [2]albertodiaz@fdi.ucm.es, [3]jmlahoz@ucm.es

Abstract

Representation of linguistic data is an issue of utmost importance when developing language resources, but the lack of a standard written form in sign languages presents a challenge. Different notation systems exist, but only SignWriting seems to have some use in the native signer community. It is, however, a difficult system to use computationally, not based on a linear sequence of characters. We present the project "VisSE", which aims to develop tools for the effective use of SignWriting in the computer. The first of these is an application which uses computer vision to interpret SignWriting, understanding the meaning of new or existing transcriptions, or even hand-written images. Two additional tools will be able to consume the result of this recognizer: first, a textual description of the features of the transcription will make it understandable for non-signers. Second, a three-dimensional avatar will be able to reproduce the configurations and movements contained within the transcription, making it understandable for signers even if not familiar with SignWriting. Additionally, the project will result in a corpus of annotated SignWriting data which will also be of use to the computational linguistics community.

Keywords: Sign Language, SignWriting, Computer Vision, Sign Language Avatar, Textual Description

1. Introduction

One of the challenges in the study of sign language linguistics is the collection and representation of linguistic data. In computational linguistics, this problem is even more crippling, since data are the basis of any computational approach to a subject.

There is an increasing interest both in society and the scientific community in sign languages, and corpora have been created for many different sign languages and with varying schemes of annotation. However, most corpora are video-based, which is equivalent to the hypothetical case of corpora of oral languages being mostly based on audio recordings.

Recordings of real utterances, both of oral or signed languages, are difficult to process computationally, whether it is for searching or managing the data, or for linguistically analyzing it and finding its structure and meaning. Video is especially difficult, since the human visual system is highly sophisticated, and emulating its processes with artificial intelligence is not a solved problem yet.

In oral languages, writing poses a useful alternative to recordings, and is indeed (and maybe to a fault) the basis on which computational linguistics have been built. However, there does not exist an equivalent in signed languages. There is not a widely accepted written form for these languages, even less a literature or a corpus of real world linguistic data that can be exploited.

There exist some candidates for this, the most promising being SignWriting. SignWriting is a system that can act as a written form of sign language, or at least as a transcription system for it. It is iconic and very in-line with the visual nature of sign languages, so it is easy to understand and accept by native signers. The problem is that it is not as easy to use in the digital world, not being formed by linear strings of characters that can be quickly input with a keyboard and consumed by the many tools developed by the computational linguistics community.

We present an early-stage project for developing tools and resources that aim to facilitate the effective use of SignWriting in computers. With these tools, input of SignWriting can be as quick as writing it on paper, and no further processing by the user is necessary. Other tools will also use this input to generate related output, such as a textual description of the signer's actions or an animated avatar, which means that SignWriting will be useful as a digital representation of sign language even for users not familiar with it. This can help in the teaching of sign language, by facilitating the use of this language in computers, and also increase accessibility and inclusion of the Deaf community in the digital world.

In the next section, we give a brief overview of the problems of sign language notation, and quickly explain SignWriting and computer vision, the artificial intelligence tool to be used for its processing. Section 3 explicates the architecture of the project and its different components, and in section 4 some conclusions are drawn.

2. Background

Sign languages are natural languages which use the visual-gestual modality instead of the oral-acoustic one. This means that instead of performing gestures with the vocal organs, which are transmitted to the receiver via sound, sign languages utilize gestures of the body, especially of the hands, to transmit information visually.

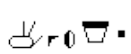
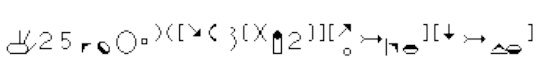
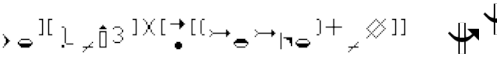
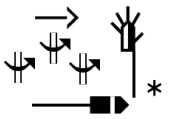
	Stokoe Notation	HamNoSys	SignWriting
Three	$3^{\perp}$		
Bears	$[]\checkmark C^{\ddagger}\checkmark C_{X}^{V\bullet}$		
Goldilocks	λY_{V}^{O}		
Deep Forest	$\overline{B}_{a}\checkmark B_{\wedge >}^{\omega}$		

Table 1: Comparison of notation systems for sign languages, using words from an American Sign Language text for the story of Goldilocks and the Three Bears[2]. The systems compared are Stokoe Notation (Stokoe, 1980), HamNoSys (Hanke, 2004) and SignWriting (Sutton, 1995).

While oral languages have developed writing systems that represent the sounds (or sometimes ideas) of the language in a visual, abstract, and standard way, none such system has organically appeared for sign languages. Writing systems have many advantages, both to users of the language in helping them analyze it, and making structure explicit, and to linguists. To linguists, one advantage of writing systems of great relevance lately, and especially to us in the computational linguistics community, is the ease of computational treatment.

A number of systems have been developed for the transcription of sign language into written form (Stokoe, 1980; Herrero, 2003; Hanke, 2004). Most of them are intended for linguistic research and transcription of fine linguistic detail, and none of them seem to have seen universal use or the kind of standardization seen in the writing systems of oral languages.

This presents a challenge for the development of language resources. Systems which are alien to native informants of sign language require training for these users, and in limited time frames inevitably pose the question of whether the information transcribed with them really is what the signer intended. Additionally, we have found that computational tools for the management of the different notation systems are not very mature or wide-spread.

However, there is another proposed transcription system for sign languages: SignWriting (Sutton, 1995). SignWriting is a system developed by Valerie Sutton, a non-linguist, in 1974, designed specifically to write sign languages. There is much information on its use and practicalities on the website[1], and especially interesting is the comparison between some notation systems[2]. We reproduce a slightly modified excerpt in table 1.

We give a short introduction to SignWriting in the following, but Di Renzo et al. (2006) give an informative discussion of the use of this system in linguistic research, along with some notes on the challenges that notation systems present. More on this topic and on the differences between notation and writing systems can be found in Van der Hulst and Channon (2010).

2.1. SignWriting

As mentioned before, SignWriting is a system intended for the writing of sign languages. It is made up of symbols, many of which are highly iconic, that represent different linguistic or paralinguistic aspects. See for example Sutton (2009).

Different handshapes (such as a closed fist, an open palm, etc.) are depicted by figures like a square, or a pentagon, respectively. Conventional strokes can be added to these basic shapes to represent the thumb or the different fingers. The spatial orientation of the hand is symbolized by a black and white color code, among other possibilities. There are also icons for different locations on the body (mainly, parts of the head and face). Other symbols stand for changes in the handshape or the orientation, for different kinds of movements and trajectories, for contacts, for variations in the speed, and for facial expressions, including eyebrow intonation and other paralinguistic realizations. Finally, there are symbols that represent pauses and prosodic grouping, thus allowing to write full sentences.

All these symbols combine non-linearly in space to transcribe signs in a visually intuitive way. This is a most welcome characteristic for the Deaf community, inasmuch as they give preeminence to anything visual, and it makes it easier to learn for students of sign languages or any interested person.

Furthermore, its iconicity, together with its flexibility, al-

[1] http://www.signwriting.org

[2] http://www.signwriting.org/forums/linguistics/ling001.html

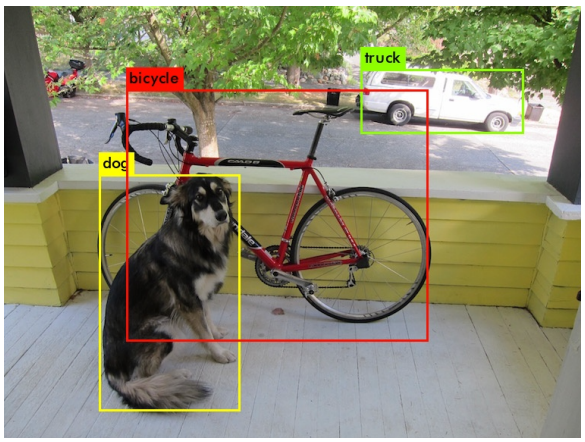


Figure 1: Object detection task, where objects in an image are located and classified (Redmon et al., 2016).

low to transcribe any newly-coined sign, making it advantageous for treating sign languages not only for daily use but also in technical, scientific, and educational environments.
The fact that symbols are not interpreted linearly, but according to their position relative to other symbols, poses a challenge to the computational treatment of these bundles. It is necessary to decompose the fully transcribed signs into their components and parametrize them in linguistically relevant subunits or features.
If SignWriting annotations are created with computer tools, this information may be readily available. However, the non-linearity of SignWriting, along with the large amount of symbols that can be used, make computational input cumbersome and far slower than hand-drawing of transcriptions. Additionally, existing transcriptions, even if computer made, may not be available in their decomposed form, but rather as a plain image with no annotation. Therefore, there exists the need for tooling that can interpret images containing SignWriting transcriptions in an automatic way.

2.2. Computer Vision

Broadly speaking, computer vision is the field of artificial intelligence where meaning is to be extracted from images using automatic procedures. What this meaning is depends on the context, the available data, and the desired result. As in other fields of artificial intelligence, classification is the task of assigning a label to an image, for example the type of object found in a photograph, or the name of the person a face belongs to.
Object detection is a step beyond, in which it is to be found in an image not only what object it represents, but also where in the image it is. In the most common case, there can be many objects in an image, or none, and it is necessary to find how many there are, where, and what their labels are.
This is a difficult task, but it is very well suited to machine learning approaches, especially neural networks and deep learning. These techniques work by presenting a large amount of annotated data to the algorithm, which is able to extract from them features and patterns from which to decide the result of the procedure. Often, this means bounding boxes: rectangles that contain the object in the image, along with labels for what the detected object is. In Fig. 1 an example of this task can be seen.
YOLO (You Only Look Once) is an algorithm for object detection that works by applying a single neural network to the full image (Redmon and Farhadi, 2018). Other algorithms work in multiple steps, for example by first performing detection of possible candidates and then classifying them, but YOLO works in a single pass, making it faster and easier to use. It works by dividing the image into regions, predicting bounding boxes and label probabilities for each region, and then collating these regions and possibilities into the final list of results. Its implementation in Darknet (Redmon, 2013 2016) is very easy to configure and utilize, while retaining precision at the state of the art.
This task of object detection is exactly what we need for understanding SignWriting transcriptions. They are formed by different symbols, placed relative to each other in a way that is meaningful and significant. By using YOLO, we can automatically find these symbols and their positions in SignWriting images, which allows us to further work with the meaning of the transcription instead of with the pixels of the image[3].

3. The VisSE project

During the authors' research in Spanish Sign Language, the problems outlined in the introduction regarding its digital treatment were patent. As students of this language as well as researchers and engineers, ideas for solutions started to come to our minds. At some point, previous expertise in image recognition, a very salient topic in sign language research, joined the knowledge of SignWriting as a useful tool for these languages, used by our educators and many in the Deaf community.
Some of the ideas for both tools and processes were combined into a single effort for which funding was requested, and granted by Indra and Fundación Universia as a grant for research on Accessible Technologies. This effort resulted in the VisSE project ("Visualizando la SignoEscritura", Spanish for "Visualizing SignWriting") aimed at developing tools for the effective use of SignWriting in computers. These tools can help with the integration of Hard of Hearing people in the digital society, and will also help accelerate sign language research by providing another methodology for its research.
A general architecture of the project can be seen in Figure 2. There, the sign in Spanish Sign Language for "teacher" is used as an example. Its transcription in SignWriting is decomposed and processed by an artificial vision algorithm, which finds the different symbols and classifies them. The labels and relative positions of the symbols are then transformed into their linguistic meaning, called here "parametrization". In the example, the usual features of sign language analysis are used, but this representation is yet to be decided, and has to follow closely the information encoded in the SignWriting transcription. The parametrization is then turned into a textual description, which allows

[3] When we say meaning of the transcription, we mean the codification it contains of sign language utterances, not the meaning of the represented signs in a linguistic semantics way.

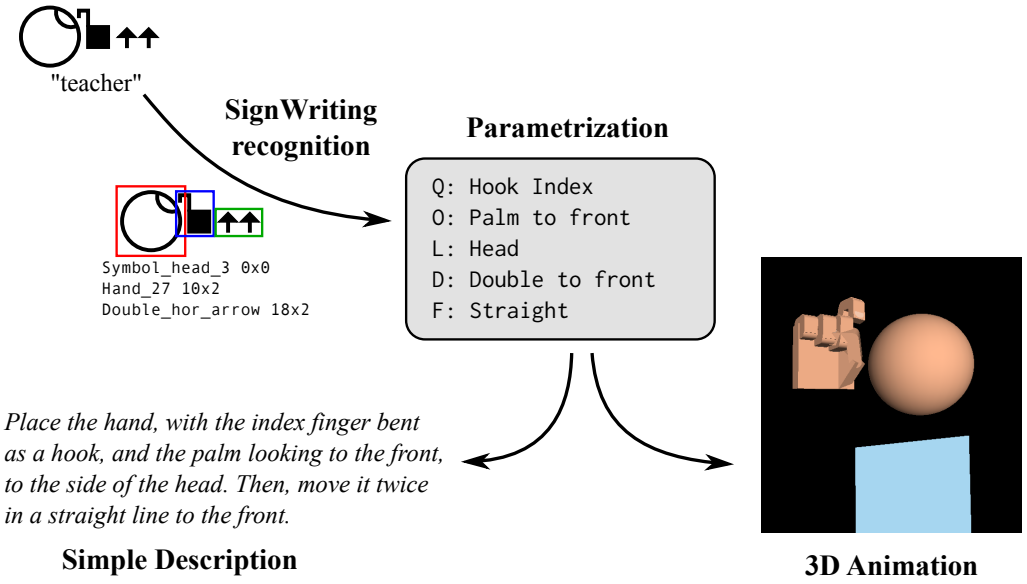


Figure 2: Architecture of the different components of the VisSE project.

a non-signer to realize the sign, and a 3D animation which can be understood by a signer.

3.1. Corpus of SignWriting Transcriptions

While the goal of the project is to develop the tools mentioned before, which will help with the use of SignWriting in the digital world, there will be an additional language resource result. Data are of paramount importance when doing computational linguistics, and the artificial vision algorithms to be used rely on these data for their successful training and use.

Therefore, one of the products of the project will be a corpus of linguistic annotations. Entries in the corpus will be, as far as possible, input by informants who are native signers of Spanish Sign Language. For this purpose, a custom computer interface will be developed. This interface needs only be a simple front-end to the database, with roles for informants and for corpus managers, and with some tool to facilitate SignWriting input, either by a point-and-click interface or by a hand-drawing or scanner technology. Annotation, however, will not consist of grammatical information, but rather of the locations and meanings of the different symbols in the transcription.

Even if less interesting to our users, this result will probably be of use to other researchers, so it too will be publicly released. Similar to other such projects, the main object of annotation will be lexical entries, words of sign language and their realization, the main difference being that the data recorded will be in the form of SignWriting. The meaning of the annotated sign will be transcribed using an appropriate translation in Spanish.

Corpora that peruse SignWriting already exist (Forster et al., 2014), and there is also SignBank[4], a collection of tools and resources related to SignWriting, including dictionaries for many sign languages around the world, and SignMaker, an interface for the creation of SignWriting images. While useful, the data available in the dictionaries are limited, especially for languages other than American Sign Language, and its interface is more oriented toward small-scale, manual research rather than large-scale, automated computation.

3.2. Transcription Recognizer

At first, the annotations in the corpus will have to be performed by humans, but they will immediately serve as training data for the YOLO algorithm explained in section 2 2 As annotation advances, so will increase the performance of the automatic recognition, which will be used to help annotators in their process by providing them with the prediction from the algorithm as a draft. This will accelerate data collection, which will in turn increase training effectiveness until at some point the algorithm will be able to recognize most input on its own.

The use of YOLO for recognition of SignWriting has already been successfully prototyped by students of ours (Sánchez Jiménez et al., 2019). The located and classified symbols found by the algorithm will then be transformed into the representation used in the corpus, which will include the linguistically relevant parameters (for example, it is relevant that the location is “at head level”, but not whether the transcription is drawn 7 pixels to the right).

This process of finding out sign language parameters using computer vision is akin to that of automatic sign language recognition in video, which is often performed for video-based corpora. However, it is much simpler, both for the human annotator and the computer vision algorithm, since images are black and white, standardized and far less noisy. Transcriptions, being composed out of a discrete (even if

[4] http://www.signbank.org/

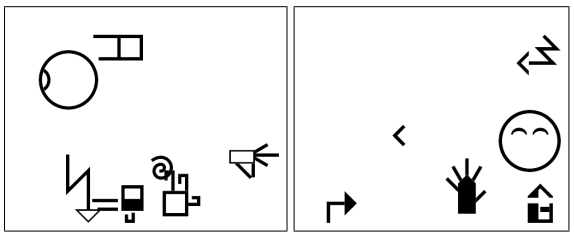


Figure 3: Automatically generated training samples for YOLO.

large) amount of symbols, present an additional advantage: training data can be immensely augmented by automatic means. An algorithm, already implemented, mixes the possible symbols to create random images which contain the data and annotations the YOLO algorithm needs to be trained. An example of this can be seen in Fig. 3. Even if these are meaningless as SignWriting transcriptions, they contain the features and patterns that the algorithm needs to learn. This will help bootstrap the system, which will make the recognizer useful to annotators earlier rather than later during the project.

3.3. Description of the Sign

Once the algorithm is able to understand transcriptions, its parametric result will be used in tools that can help integrate the Deaf community in Spain into the digital world.

The first such tool will be a generator of textual descriptions from SignWriting parametrizations. This description will explain, in a language easy to understand, the articulation and movements codified in the SignWriting symbols.

This will be a useful aid in communication between signers and non-signers. For example, it can be explained to non-Deaf people how to sign basic vocabulary, like pleasantries, or maybe important words in a particular domain (an office, a factory). This will allow non-Deaf people communicate to Deaf people basic information, useful for the daily routine or maybe an emergency, without the need to learn sign language proper. This may help broaden the employability landscape for Deaf people, increasing their inclusion in the office community and with their non-signing peers.

The use of text instead of video has some advantages. While observation of real video and images is necessary for proper understanding of the rhythm and cadence of sign language, it is often not enough for correct articulation and orientation of the hand. Non-signers are not used to looking for the visual cues in hand articulation, and may confuse the hand configurations necessary for particular signs. Spelling them out, however, can help them realize the correct finger flexing and wrist rotation, in an environment where an interpreter or teacher may not be readily available.

This also leads to a different application of this tool: education. Both Deaf and non-Deaf people can find it challenging to self-study sign language, due to the scarcity of resources and the challenge of a lack of a linear transcription system. Translating SignWriting into text can improve understanding of this system, helping both signers and students learn the representation of sign language in SignWriting symbols in a dynamic environment with immediate feedback.

3.4. Animated Execution by an Avatar

Another result of the project will be a three-dimensional animated avatar, capable of executing the signs contained in the parametric representation. SignWriting is an (almost) complete transcription of sign language, including spatial and movement-related information. This information, after its computational transformation into parameters, can be directly fed into a virtual avatar to realize the signs in three dimensions.

The advantages of the use of avatars are known in the community, and have been studied before. Kipp et al. (2011) give an informative account of different avatar technologies and the challenges in their use, and propose methodologies for testing and evaluation of the results. Bouzid and Jemni (2014) describe an approach very similar to ours in its goal, using SignWriting as the basis for generating the sign language animations. However, this process is not done automatically, but manually as in other avatar technologies.

Manual preparation of the execution of signs is a costly procedure, even if not as much as video-taping interpreters. An expert in the system as well as one in sign language are needed, and it is difficult to find both in one person. With our approach, instead of knowing the intricacies of the avatar technology, just an expert in SignWriting would be needed. It is far easier that this expert is the same as the sign language translator or author, or minimal training can be provided. SignWriting is also easier to carry around and edit, compared to systems like SIGML (Kennaway, 2004), which may be intuitive and easy for computer engineers but no so much to non-computer-savvy users.

Our system will also strive to be dynamic, not presenting a static sequence of images but rather an actual animation of the sign. While sign language generation is very complicated, it is important to note that this is not what our system needs to do. Placement and movement are already encoded in SignWriting, and our system only needs to convert the parameters into actual coordinates in three-dimensional space. The technology for this tool will be Javascript and WebGL, which are increasingly mature and seeing wide-spread adoption in the industry. Web technology is ubiquitous nowadays, and browsers present an ideal execution environment where users need not install specific libraries or software but rather use the same program they use in their everyday digital lives.

4. Conclusions

As we have seen, the goal of the VisSE project is to develop a number of tools for the use of SignWriting effectively in computers. First, the recognizer will allow SignWriting to be used as input in a comfortable way. Users will not need to search for symbols, and then drag and drop them to the canvas, nor will they have to memorize an arbitrary mapping from ASCII characters to SignWriting symbols. Hand written transcriptions or existing images will be able to be processed, making the use of SignWriting practical for continuous use. Then, the description generator and the avatar will use this input to transform SignWriting transcriptions of signs into alternative representations, which will help users understand both the meaning and the use of SignWriting.

All the developed tools will be publicly released, and the full pipeline might include software that allows a user to dynamically input SignWriting into an interface and immediately watch its realization by the avatar. The data generated in the form of the corpus can also be transformed into a dictionary, one where words are stored and indexed directly in sign language. Often, sign language resources are only accessible via oral language glosses, but the use of SignWriting allows sign language to be the primary language in its own dictionary.

These are all future works worthy of research and development, which will benefit the Deaf community in Spain. But the methodology and principles used are not specific to Spanish Sign Language, so we expect they will be able to be adapted to other sign languages.

Apart from the results benefiting the Deaf community, there will also be results for the language resource community. The data collected, in the form of the corpus, and the recognizer algorithm, will be released for the use of other researchers. Additionally, if this project helps SignWriting to become even more widespread and easier to use in computational contexts, this might become another powerful tool for the sign language linguistics community.

Therefore, we present this article to the community, with the goal of receiving feedback and comments during the early stages of the project so that it can inform and improve its development and its usefulness for the Computational Linguistics field.

5. Acknowledgements

The research leading to and contained within this project is partially funded by the project IDiLyCo: Digital Inclusion, Language and Communication, Grant. No. TIN2015-66655-R (MINECO/FEDER) and the FEI-EU-17-23 project InViTAR-IA: Infraestructuras para la Visibilización, Integración y Transferencia de Aplicaciones y Resultados de Inteligencia Artificial (Universidad Complutense de Madrid).

Funding for the development of the project "Visualizando la SignoEscritura" has been awarded by Indra and Fundación Universia as part of the program for funding of research projects on Accessible Technologies.

6. References

Bouzid, Y. and Jemni, M. (2014). A virtual signer to interpret SignWriting. In Klaus Miesenberger, et al., editors, *Computers Helping People with Special Needs*, volume 8548, pages 458–465. Springer International Publishing.

Di Renzo, A., Lamano, L., Lucioli, T., Pennacchi, B., and Ponzo, L. (2006). Italian Sign Language (LIS): can we write it and transcribe it with SignWriting? *Proceedings of the 2nd Workshop on the Representation and processing of Sign Languages "Lexicografic matters and didactic scenarios" – LREC*, pages 11–16.

Forster, J., Schmidt, C., Koller, O., Bellgardt, M., and Ney, H. (2014). Extensions of the sign language recognition and translation corpus rwth-phoenix-weather. In *LREC*, pages 1911–1916.

Hanke, T. (2004). HamNoSys – Representing Sign Language Data in Language Resources and Language Processing Contexts. In *Proceedings of the Workshop on Representation and Processing of Sign Language, Workshop to the forth International Conference on Language Resources and Evaluation (LREC'04)*. ISSN: 17913721.

Herrero, Á. (2003). Escritura alfabética de la lengua de signos española: once lecciones. *Escritura alfabética de la lengua de signos española*, pages 1–159.

Kennaway, R. (2004). Experience with and Requirements for a Gesture Description Language for Synthetic Animation. In Gerhard Goos, et al., editors, *Gesture-Based Communication in Human-Computer Interaction*, volume 2915, pages 300–311. Springer Berlin Heidelberg, Berlin, Heidelberg.

Kipp, M., Heloir, A., and Nguyen, Q. (2011). Sign language avatars: Animation and comprehensibility. In Hannes Högni Vilhjálmsson, et al., editors, *Intelligent Virtual Agents*, volume 6895, pages 113–126. Springer Berlin Heidelberg.

Redmon, J. and Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv*.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788.

Redmon, J. (2013–2016). Darknet: Open source neural networks in c. `http://pjreddie.com/darknet/`.

Sánchez Jiménez, J. B., López Prieto, S., and Garrido Montoya, J. Á. (2019). Reconocimiento de lenguaje signo-escritura mediante deep learning.

Stokoe, W. C. (1980). Sign language structure. *Annual Review of Anthropology*, 9(1):365–390.

Sutton, V. (1995). *Lessons in sign writing*. SignWriting.

Sutton, V. (2009). Signwriting: sign languages are written languages. *Center for Sutton Movement Writing, CSMW, Tech. Rep*.

Van der Hulst, H. and Channon, R. (2010). Notation systems. In Diane Brentari, editor, *Sign languages*, pages 151–172. Cambridge University Press, Cambridge.

Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 209–216
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

Video-to-HamNoSys Automated Annotation System

Victor Skobov, Yves Lepage
Graduate School of Information, Production and Systems
Waseda University
v.skobov@fuji.waseda.jp, yves.lepage@waseda.jp

Abstract

The Hamburg Notation System (HamNoSys) was developed for movement annotation of any sign language (SL) and can be used to produce signing animations for a virtual avatar with the JASigning platform. This provides the potential to use HamNoSys, i.e., strings of characters, as a representation of an SL corpus instead of video material. Processing strings of characters instead of images can significantly contribute to sign language research. However, the complexity of HamNoSys makes it difficult to annotate without a lot of time and effort. Therefore annotation has to be automatized. This work proposes a conceptually new approach to this problem. It includes a new tree representation of the HamNoSys grammar that serves as a basis for the generation of grammatical training data and classification of complex movements using machine learning. Our automatic annotation system relies on HamNoSys grammar structure and can potentially be used on already existing SL corpora. It is retrainable for specific settings such as camera angles, speed, and gestures. Our approach is conceptually different from other SL recognition solutions and offers a developed methodology for future research.

Keywords: sign language, machine learning, HamNoSys, corpus annotation

1. Introduction

The Hamburg Notation System 2.0 (HamNoSys) was presented by Prillwitz et al. (1989). An updated version 4.0 is described in (Hanke, 2004). It is meant to annotate movements, using a set of approximately 200 symbols and a standardized structure. This notation system is capable of transcribing any sign in any sign language (SL), which makes it a significant contribution to the current state of sign language research, not only because of its utility in corpus annotation but also because of its ease of processing. HamNoSys has a Unicode-based implementation[1] and has a fully developed markup language implementation Sign Gestural Markup Language (SiGML).

In this paper, by HamNoSys, we will be referring to its manual markup language representation - SiGML, as described by Elliott et al. (2004). In particular, SiGMLs manual features descriptions, the part that implements HamNoSys symbols as inner elements. SiGML can also include non-manual features of sign: pose, eye gaze, mouth, and facial expressions. The production of virtual avatar animations from SiGML was first presented by Kennaway (2002) and is used in the JASigning[2] platform.

The vast majority of sign language data is presented in a video-based corpus format and includes signs labeled along the timeline. Some data already include HamNoSys annotations (Hanke, 2006). However, not all available SL corpora are annotated with HamNoSys because such annotation is time-consuming. In particular, it requires a good understanding of the HamNoSys grammar. Having SL corpora annotated in HamNoSys would simplify and ease the cross-language research and SL data processing. The Dicta-Sign Project (Matthes et al., 2012) aimed at this by creating a parallel multilingual corpus. Their work provided a set of tools for direct HamNoSys annotation and allowed us to synchronize the generated animations with video data. Efthimiou et al. (2012) showed how to use image processing to find a matching sign from the vocabulary automatically, which improved annotation speed. However, video and image data require heavy processing. For example, Östling et al. (2018) presented comprehensive research on 31 sign languages and 120,000 sign videos. The location and movement information was collected using video processing. The collection alone took two months of computing time on a single GPU. Another problem with video-based corpora is an issue with signers' privacy. Most of the SL corpora that available today for researchers include a signing person. Which often leads to limited accessibility of the corpora.

Dreuw et al. (2010) presented a sign language translation system as a part of The SignSpeak Project, which includes an automatic sign language recognition system that is based on sign feature extraction from tracking the body parts on video frames. Curiel Diaz and Collet (2013) proposed another semi-automatic sign language recognition system, which also relies on head and hands tracking on video frames and uses a Propositional Dynamic Logic.

Still, many SL corpora do not include HamNoSys annotations. Hrúz et al. (2011) used an existing dictionary with annotated video data and custom classes for the categorization of signs in sign language corpora, making substantial progress towards annotating new data. However, such methods rely on an existing vocabulary, which is not always available. Ong and Ranganath (2005) discussed how a large-vocabulary recognition system still has to deal with sign translation and cannot fully grasp the actual meaning behind the signs and their different lexical forms. Additionally, the authors investigate how the annotation of signs allows us only to a certain degree to have an understanding of their meaning. Ebling et al. (2012) mentioned the problem of expressing the vocabulary of sign language using a spoken language. In order to have a raw sign language without interpretation and have full use of character-based natural language processing methods, an automated sign movement annotation system is required.

[1]HamNoSysUnicode: `http://vhg.cmp.uea.ac.uk/tech/hamnosys/HamNoSysFonts.pdf`

[2]JASigning: `http://vh.cmp.uea.ac.uk/index.php/JASigning/`

2. Methodology

The main goal of our research is to develop a Video-to-HamNoSys decoder, having a HamNoSys-to-Video encoder. Thus, we propose our system in the form of the Encoder-Decoder model demonstrated in Figure 1.

For the training of the automatic annotation system, we require an extensive data set of sign videos, correctly transcribed in HamNoSys. The system should be able to transcribe any sign movement. Consequently, the training dataset has to be diverse and language-agnostic. To this end, we need a generation method of random HamNoSys annotations that incorporates all rules of the annotation system. Each generated sign annotation has to be accepted by the JASigning virtual avatar animation tool to produce signing animations. We use virtual avatar animations of the JASigning platform to synthesize movements from generated annotations. To cover all possible HamNoSys movements for any sign language, we need to generate signs randomly, representing random and chaotical movement. The encoding process will be handled mostly by JASigning software, and It will be detailed in Section 3..

The preprocessing steps, described in Section 4.1., will allow us to generate the necessary training data to train the decoder. The decoding process will be taken care of by our proposed tree-based machine learning model, detailed in Section 4..

3. Encoder

In this section, we first introduce a new generation method with a grammar tree that plays an essential role in the whole system. We describe the building process and the properties of the generation tree in the next Subsection, 3.1.Generation grammar tree. We then introduce the generation method with the generated data set that will be analyzed and compared to existing HamNoSys data. The overview of the encoder is shown in the upper part of Figure 1.

3.1. Generation Grammar Tree

To build the generation grammar tree, we start with the basic HamNoSys structure presented in the upper part[3] of Figure 2, and worked through each of the single rule blocks: Handshape, Hand position, Location, and Action. The notation for two hands was added afterwards. The verification was done by running the JASigning avatar animation (Kennaway, 2002). The process of adapting the rules was as follows: first, add the simplest ones; then add the specific rule cases. The presented work by Hanke (2004) was used as a description and guide of the HamNoSys. For the grammar source, we utilized the SiGML Data Type Definition (DTD) *hamnosysml11.dtd*[4] files with regular expressions. In particular, the *hamnosysml11.dtd* file provides a good overview and a basic understanding of HamNoSys grammar.

The upper part of Figure 2 displays the basic HamNoSys structure for one hand annotation. The lower part of Figure

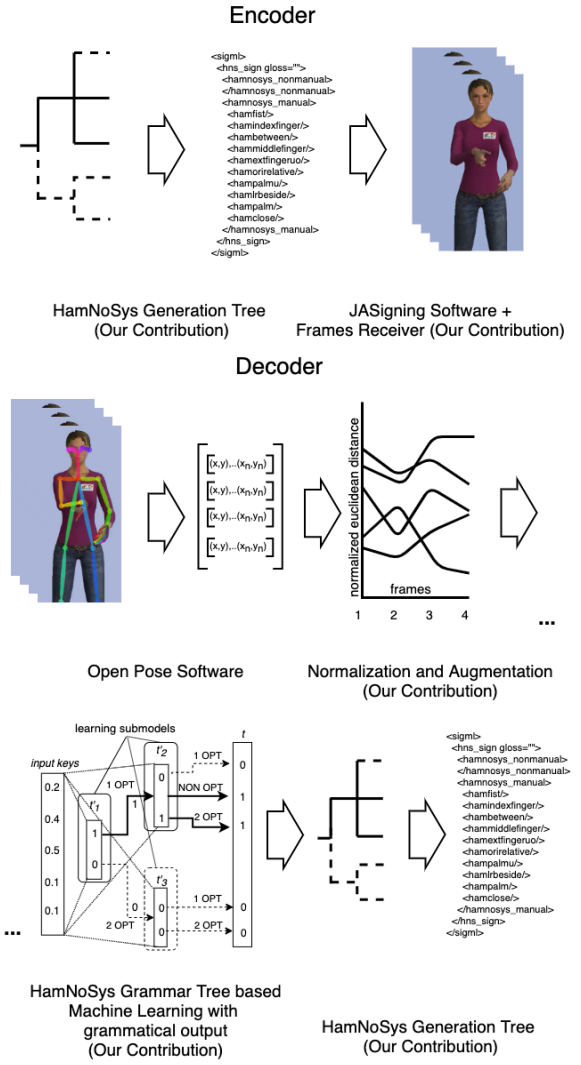


Figure 1: Representation of a tree-based learning system: the upper part shows the encoder, the middle and the lower parts show the decoder of the system

2 is a representation of the same rule in the form of a tree. The root node is a non-terminal grammar symbol. Leaf nodes are representing the rule members and their order. Between the root and the leaves, there are production nodes, which are named "#_OPT" for optional ("#" stands for the number of the option) or "NON_OPT" for non-optional, i.e., obligatory nodes. The optional and non-optional nodes are necessary for the generation algorithm, which will be described in Subsection 3.2.. The lower part of Figure 2 shows the tree leaves which bear *empty*, representing an empty string or symbol in the grammar.

3.1.1. Regular Expressions

Figure 3 shows the examples of regular expressions implemented in a tree form. To describe a boolean *or* in the grammar, the "NON_OPT" node was used with "#_OPT" children for each *or* case. For *zero or one*, we used two "#_OPT" nodes, one of them having an "empty" node as a

[3] Available at https://robertsmithresearch.files.wordpress.com/2012/10/hamnosys-user-guide-rs-draft-v3-0.pdf

[4] Available at http://www.visicast.cmp.uea.ac.uk/sigml/

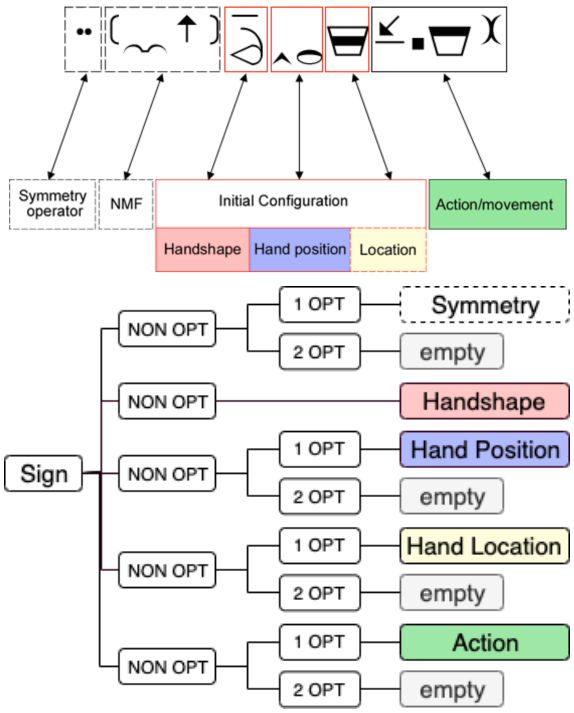


Figure 2: Implementation of the HamNoSys general structure (Up) in tree form (Down).

child. For *zero or more* and *one or more* we used optional "#_OPT" nodes. Any loop or recursive listing is limited only by **two** mentions. For example, the notation of different actions one after another could be infinite, but in order to introduce such rule of repeated notations of actions, and avoid infinite loops, the amount for actions was limited to a maximum of **two** actions.

3.1.2. Symbol Order

The HamNoSys system is order specific. Tree representation of a rule has to respect the order of symbols in the grammar. Keeping the order of leaves in the tree, according to the HamNoSys grammar, is crucial. In the next Subsection 3.2., the generation algorithm returns leaves in a post-order traversal of the leaves, allowing the algorithm to keep the grammatical order of the returned terminals.

3.1.3. Individual HamNoSys Sign Parts

The tree building process continues from root to the terminals. It is possible to build a generation tree for a single HamNoSys non-terminal element for elements like handshape, location, movement, etc. This might be useful if a change in a part of sign notation is needed. The generation of single sign parts was used for validation and modification of a single rule, during the conversion process.

3.1.4. Tree Limitations

During the process of building a sign generation tree, we had to remove a number of HamNoSys symbols, listed in Table 1. Most of them were excluded because they represent sign sentence and text markers: punctuation and location pointers. One symbol *hamupperarm* was removed

hamexclaim	*hamcomma*	*hamspace*
hamfullstop	*hamquery*	*return*
hammetaalt	*hamaltend*	*hamnbs*
hamcorefref	*hamupperarm*	*hametc*
hamaltbegin	*hammime*	*tab*
hamcoreftag	*pagebreak*	*linefeed*
hamnomotion	*hamversion40*	

Table 1: 20 HamNoSys symbols removed from the generation tree

from the hand *location1* rule, because the parser did not accept it. It is also worth mentioning that the *replacement* rule produces a visualization error during the animation, but it was kept regardless. By expanding the HamNoSys grammar, the generation tree can be recreated and updated.

3.1.5. Tree Format

For each of the 83 extracted rules, a tree representation was created with a non-terminal element as its root. These representations were combined into one large complete generation tree, with HamNoSys terminal symbols or *empty* symbols as leaves. The generation tree contains 302,380 leaves and stored in *Newick* format, (.nk or .nwk file). This format can store internal node names, branch distances, and additional information about each node encoded as features. For tree processing, we used the ETE Toolkit[5] package for *Python 3.6*. The resulting tree can produce any valid sign in any sign language, with only the limitations mentioned above. The 83 extracted rules and the generation tree will be released publicly with this paper, and open for future improvement.

3.2. Sign Generation Process

Having a grammar tree structure with all terminal elements as leaves and all non-terminals and internal nodes allows us to retrieve HamNoSys symbols as leaves. This can be achieved by traversing the tree and returning only the required symbols. We first mark the needed symbols. Marking is provided according to the grammar, i.e., the tree structure itself. We developed Algorithm 1 for this process. By using a top-down traversal, from the root to the leaves, the algorithm uses the optional "#_OPT" nodes to select a path to the leaves, which allows us to retrieve and mark the required terminal symbols.

The "#_OPT" optional nodes are used to introduce randomness into the generation process. In Algorithm 1, the goal is to generate a random sign, which represents a random movement. We create a data set of signs distributed over the whole HamNoSys grammar. We can modify the generation process by assigning weights to optional nodes. We can encode multiple signs on the generation tree with probability values between 0 and 1.

The HamNoSys notation allows a precise description of movements: specific annotation for each finger, complex location and position annotation of hand, repetitions, and symmetry of hand movements. The more precise the description, the more complicated the rule to be applied, and

[5] Available at `http://etetoolkit.org/`

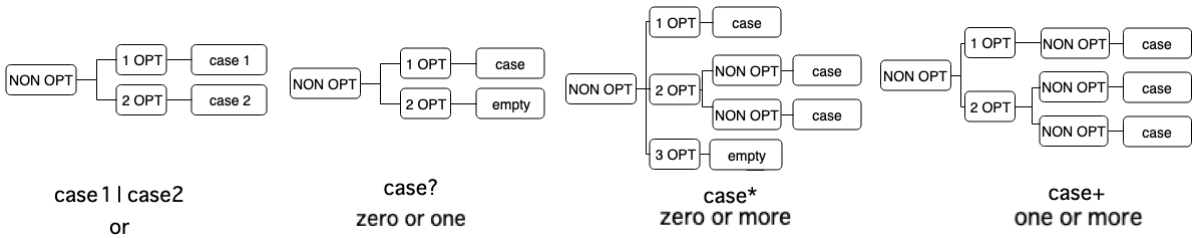


Figure 3: Implementation of regular expressions in tree form

Algorithm 1 Recursive tree walker for weighted generation

```
1: function SIGNGENERATION(treenode)
2:     if tree node is leaf then                          ▷ Check whether the node is a leaf
3:         return tree node                               ▷ Terminal retrieved
4:     else
5:         Initialize C ← ∅
6:         for all child ∈ tree node.children do
7:             if child.name ='OPT' and ≠ 'NON OPT' then
8:                 C,p_c ← child ∪ C                      ▷ Getting set of optional children nodes and their probabilities
9:         if C ≠ ∅ then                                  ▷ Pick one child from the set according to their probabilities
10:            Initialize randomChild ← WeightedRandomFromSet(C,p_c)
11:            for all child ∈ tree node.children do
12:                if child.name ='NON OPT' or child = randomChild then
13:                    SignGeneration(child)              ▷ All non optional and one picked optional child
14:        else
15:            for all child ∈ tree node.children do
16:                SignGeneration(child)                  ▷ If option Set is empty, go further to each child
```

the more complicated the rule, the longer the tree branch. Due to the differences in the topology distances between the root node and leaves, a random selection of optional nodes will produce an unequally distributed set of terminal symbols. As a result, the leaves with longer branches will occur less frequently in the generated sign. The use of such a data set may lead to over- or under-classification problems. Our goal is to provide a source of HamNoSys data, which will be used to train a machine learning system. In the ideal case, all leaves must have equal probabilities of being included in the generated sign, in a balanced data set. Consequently, equalization through the weighted selection process of optional nodes is required.

To accomplish this, in Algorithm 1: before selecting a random child on line 10, we compute the weights of all "#_OPT" optional nodes, according to their probabilities given by Formula 1. For a "#_OPT" child node, we take into account the sum of all leaves under all optional sister nodes and calculate the rate of leaves under each optional node. The probability value stored as a branch distance to the parent node due to its accessibility in the Newick format. With all rates computed, we perform a weighted random selection (WRS), which is provided as function *random.choices(weights)* in the package *random*[6] of *Python3.6* programming language to perform this.

$$p_{opt} = \frac{opt.leavesnumber}{\sum_{opt \in \mathcal{C}} opt.leavesnumber} \quad (1)$$

We discuss the result of leaf probability equalization and its effect on the generated data set in the next section.

3.3. Generated Data

Using the proposed tree structure, we were able to generate data set with 10,000 signs with the weighted random selection (WRS) approach. The generation of 1 sign takes approximately 3.6 sec[7]. Figure 4 shows the distribution of generation tree leaf appearances in the generated sets. Due to the large size of the generation tree, it is hard to show a leaf with zero occurrence in the figure. Table 2 "Percentage of unused leaves" column gives the percentage of leaves with zero occurrences in the set. The desired data set should include all leaves to represent all grammar features of HamNoSys.

3.3.1. Weighted Random Selection (WRS) Data Set

Table 2 indicates that 13 % of the leaves did not appear in any sign of the WRS data set. The average single sign length is 357.98 symbols, and the longest sign consists of 508 symbols, which demonstrates the complexity of HamNoSys and its capability for describing complex movements.

[6]Description at: `https://docs.python.org/3.6/library/random.html#random.choices`

[7]on a single machine with Intel i7 processor 16 GB of RAM using CPU only

Source	Unique Signs	Single Sign Length Stats			Percentage
		Minimal	Average	Maximal	of unused leaves
Open DGS-Corpus	5,475	1	13.02	75	-
Generation with WRS	10,000	113	357.98	508	**12.73**%

Table 2: Comparison of signing data sets and their features

To compare the generated signs with existing natural language signs, we use the Open German Sign Language Corpus (DGS-Corpus)[8] presented by Prillwitz et al. (2008). Figure 4 shows that the average sign length for the DGS set is 13.02, which is significantly lower than in WRS generated set. It means that the DGS rule complexity is "covered" by the generated data.
The generation program written in *Python3.6* for generating the data will be released publicly with this paper. The WRS generation can be used as an example of encoding the tree weights and generation of signs with specific features.

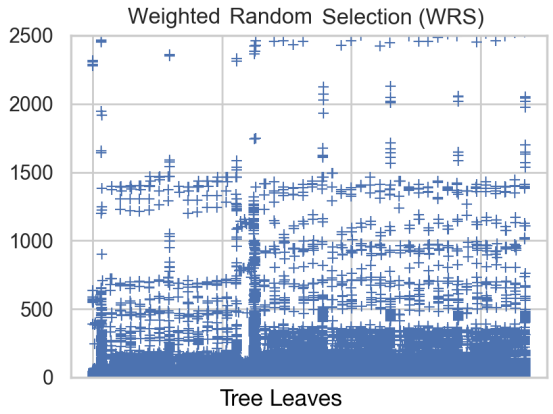


Figure 4: Impact on the distribution of 302,380 tree leaves occurrences in the generated sign set. 10,000 signs using a weighted random selection (WRS) algorithm between optional nodes.

3.4. Animation and Video Frames

The future system must be able to transcribe video materials, so it will receive video frames as an input and HamNoSys annotation as a SiGML file (.sigml) as an output. For that reason, the video of the generated SiGML signs has to be created. With the help of the JASigning tool, it is possible to produce animations virtual avatar from SiGML notation and send video frames of the animation on the network port. An example of the animation frame is demonstrated in Figure 5 (Left).

4. Decoder

In this section, we propose the Video-to-HamNoSys decoder as a part of our automatic video annotation system. The overview of the decoder is shown in the lower part of Figure 1. There are differences in camera settings, angles, and positions relative to the signer across different sign language video corpora. In Section 3.4., we showed how to produce animation frames with the JASigning software and store them as images. During this process, we define the camera angle and position in the virtual space. This step can be adjusted to match the camera setting of existing video corpora.

4.1. Data Preparation

Video frames are needed as training data for our machine learning model. To output a movement transcription as a string of characters from a given video, we do not need to input all the content of video frames, but only that information, which is required for the transcription. For that, we preprocess the video frames and extract the necessary significant movement features. The data preparation process consists of the following steps: animation, keypoints extraction, normalization, augmentation. Each of these steps is detailed below.

4.1.1. Body Keypoints Extraction

After getting the animation frames for each generated sign, we use the OpenPose Software (Cao et al., 2018) to extract the body keypoints. OpenPose can detect 18 points on the body, 25 points on each hand, and 60 points on the face (Simon et al., 2017). Instead of using convolutional layers, we use a pre-trained OpenPose model to extract significant features from the frames. This approach reduces the size of training data and ensures that our decoder model receives only the necessary information about the body movement. The right part of Figure 5 shows the keypoints detected by OpenPose in the frame shown on the left part of Figure 5.

4.1.2. Normalization and Augmentation

OpenPose extracts the coordinates of the body keypoints from frames. The real camera settings in existing SL corpora can be very different from the virtual camera settings in the JASigning software. For instance, different resolution and aspect ratio may be used. Therefore, we added a normalization step to the decoder that should detach the body keypoints from the frame coordinates, and represent them as an array of distances to each other. The unnecessary keypoints, like legs and hips, are removed because they are not involved in the signing process. We calculate the euclidian distances between the remaining keypoints and normalize them against the *base body distance*. The distance that is not significantly changing during the signing process should be set as a *base body distance*. In our case, it is the distance between two shoulder points. Its value is set to 1, and the rest of the calculated distances are put into relation to it. As a result of the normalization step for each frame, all of the sign data is represented by one matrix, which stands

[8]HamNoSys was extracted from i-Lex files Available at https://www.sign-lang.uni-hamburg.de/dgs-korpus/index.php/welcome.html

Figure 5: Example of body keypoints (on the right) extracted from a single animation frame (on the left)

for the sign movement. On some of the frames, when the keypoint detection fails, we keep the default value without any change.

As an example, in Figure 6, the movement of one hand getting closer towards the head is represented by decreasing curves, which stands for the distances between the hand and the head keypoints. As the synthetic avatar movements differ from real human movement, and so as to adapt our model to real environment, with the aim to perform well in annotating real human signs, we augment the avatar data with noise.

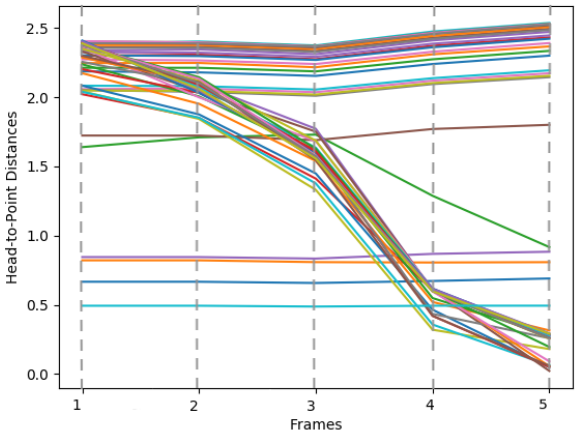


Figure 6: Plotted changes in body keypoint distances to the head keypoint across five frames

4.2. Tree-based Machine Learning System

The resulting decoder has to approximate the whole HamNoSys grammar, to generate grammatically correct output. This can theoretically be achieved by training a deep neural network, on a large number of data. We believe that such an approach is possible today, although it is not transparent. There is a risk that the output will be ungrammatical. This would also require high computational power and resources.

We suggest an alternative way. Since we have all grammar rules at our disposal in one united tree structure (that we used for the generation of the training data), we can exploit this structure to create a learning model based on the large grammar tree, where each output sign will be generated with regard to the extracted body keypoints of the initial animation movement presented in Section 4.2.1. and illustrated by Figure 7. Similar to the generation process, by traversing the generation tree, we will initialize a small learning submodel on the nodes that have optional "#_OPT" nodes as children. (Given in Section 3.2.) Algorithm 1 is used in the training process, but instead of selecting the optional "#_OPT" nodes randomly with WRS, the leaning submodel learns the right selection of optional "#_OPT" nodes. The whole generation path is learned by a chain of submodels based on the sign movement.

After the sign has been generated, it is represented by a one-dimensional vector of the generation tree's leaves. If the leaf is included, it is labeled as 1, 0 if not. This allows us to trace back each rule node in the generation tree, and see the sign recognition task as a rule-classification problem. By visiting the same leaves as during the initial sign generation, the Decoder has to repeat the initial generation process done by the Encoder. Nodes and their associated learning submodels that were not visited during the initial sign generation will not be visited during training, and their submodels will not be trained. In this way, it is guaranteed that the resulting system will always output a grammatically correct HamNoSys annotation.

Figure 7 shows the simplified underlying representation of this idea. Submodels t'_1 and t'_2 are initialized based on the main target vector t. All non-optional nodes are not learned and are set to 1 automatically. As shown in Figure 7, the node is not visited and its submodel t'_3 is not trained, since it is not involved in the generation process of the target sign at hand. This prevents oversampling on negative examples. However, the drawbacks are a reduction of training samples on the deeper levels of the generation tree and, consequently, a degradation of the accuracy, described in Section 5..

4.2.1. Learning Submodels

Each submodel has a dropout layer with a probability of 10 % (p=*0.1*), and a fully-connected layer with a LeakyRelu and Softmax activation functions. As a loss function, The Cross Entropy Loss was used, with the Adam optimizer (Kingma and Ba, 2015). The size of the output is different among all submodels, and it corresponds to the number of optional "#_OPT" children nodes, also viewed as subclasses of the tree-based machine learning system. During the training of a submodel, a subtarget vector is produced, based on the main target. It indicates whether the leaf of the corresponding optional node is included in the target sign. As shown in Figure 7, t'_1 elements are computed from t.

During tree traversal, we preprocess all the training data

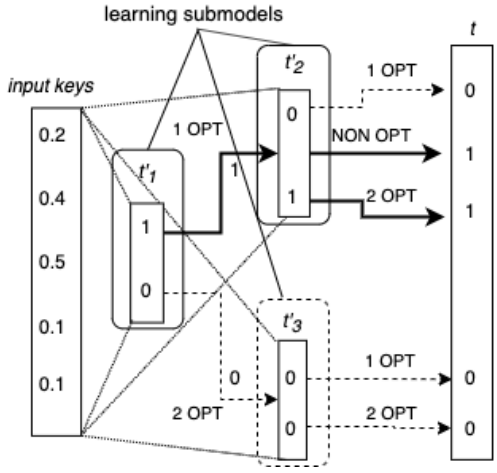


Figure 7: Representation of a tree-based learning system

and extract the subtargets to create training data subsets for each submodel. Table 3 show the number of samples that is possible to retrieve for submodels on different tree levels. We used this approach to analyze the accuracy of submodels on a different tree-levels; this is detailed further in Section 5..

Tree Level	SM	Avg. N per SC	Avg. SC	Avg. Valid. Accuracy
1	1	4561	2	93 %
2	2	1701	3	80 %
3	1	222	12	31 %
4	22	89	2	47 %
5	62	237	4	42 %
6	474	113	2	46 %
7	332	86	4	32 %
8	268	35	2	51 %
9	496	33	4	33 %
All Levels	1658	81	3	**40 %**

Table 3: SM - Number of Submodels; N - Number of training samples; SC - Number of Subclasses; Training efficiency on diferent levels of learning tree

5. Experimental Results

For our experiments, we used a part of the grammar tree, that stands for the description of handshapes. In the upper part of Figure 2, the handshape description is marked with a red color. We generated 60,000 handshape signs for one hand and extracted the body keypoints from them with the same camera position. That took us nearly ten days. The generation of the training data could potentially be simplified if JASigning Software would allow extracting the body keypoints directly from the animation.

Training of all 1,658 submodels with 800 epochs each took around five days. After implementing multiprocessing and utilizing 11 threads, we managed to reduce this time to 30 hours.

Training the tree-based machine learning system resulted in 22 % accuracy on a validation set of the 1,000 sign. In our words, our system correctly predicted 5,728 sized vector that represents a sign. The average accuracy among all submodels in the tree-based machine learning system is 40 % with an average of three classes, and an average of 81 samples per class. It is important to notice that the validation set was generated with a Weighted Random Selection (WRS) Algorithm 1 like the training data.

Our proposal being conceptually new and fundamentally different; it is not directly comparable to other approaches. Table 4 is a comparison attempt, where we give the number of subclasses in our model against the number of signs in other approaches. Our accuracy of 54 % seems to be significantly lower than the results of other models, but the reader should notice that the number of classes is much bigger.

Further investigation of our results led to an interesting finding: the degradation of the accuracy on the deeper levels of the learning tree. Table 3 gives a comparison of the average accuracy of the submodels on the different tree levels, and the amount of training samples with the average number of subclasses. Typically, for machine learning classification algorithms, the accuracy drops for a lower number of training samples and a higher number of classes. Consequently, our results might improve if we increase the number of training samples.

Image Classifier	Unique Classes	Accuracy
(Bheda and Radpour, 2017) Deep CNN	32	82 %
(Bantupalli and Xie, 2018) Deep CNN	100	93 %
Body Keypoints Classifier		
Our Model 1 Frame Average Submodel Accuracy	3	38 %
across all Subclasses	**5,728**	**9 %**
Our Model 5 Frames Average Submodel Accuracy	3	40 %
across all Subclasses	**5,728**	**22 %**

Table 4: Comparison to other arproaches

6. Discussion and Future Work

We presented an automatic annotation system that relies on HamNoSys grammar structure. Proposed approach can potentially be used for annotation of any hand movement.

For future work, we suggest modifying the training data generation process, by extracting the body keypoints from the JASigning virtual avatar directly, skipping saving the frames, and processing them with OpenPose Software. That will allow the generation of the comprehensive training data set, with the whole HamNoSys grammar included, and different camera angles. Potentially generation process could be done during the training "on the fly."

Modifications to data normalization and augmentation can be done, to achieve better performance on the real data. The model should learn the differences between the movements of an avatar and multiple human signers. Additionally, the automatic annotation system for facial expressions

and other non-manual features by further use of SiGML attributes will expand the system utilization.

By representing an entire corpus with HamNoSys strings of characters and JASigning generated animations, the size of the corpus could drastically decrease. In this way, processing and sharing corpus data might become more accessible. Processing of SL corpora as strings of characters allows the use of sophisticated NPL techniques for sign language analysis and research. This opens new directions in SL research.

7. Bibliographical References

Bantupalli, K. and Xie, Y. (2018). American sign language recognition using deep learning and computer vision. In Naoki Abe, et al., editors, *IEEE International Conference on Big Data, Big Data 2018, Seattle, WA, USA, December 10-13, 2018*, pages 4896–4899. IEEE.

Bheda, V. and Radpour, D. (2017). Using deep convolutional networks for gesture recognition in american sign language.

Cao, Z., Hidalgo, G., Simon, T., Wei, S., and Sheikh, Y. (2018). Openpose: Realtime multi-person 2d pose estimation using part affinity fields.

Curiel Diaz, A. T. and Collet, C. (2013). Sign language lexical recognition with Propositional Dynamic Logic. In *51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages pp. 328–333.

Dreuw, P., Ney, H., Martinez, G., Crasborn, O., Piater, J., Moya, J. M., and Wheatley, M. (2010). The signspeak project - bridging the gap between signers and speakers. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Ebling, S., Tissi, K., and Volk, M. (2012). Semi-automatic annotation of semantic relations in a swiss german sign language lexicon. In *5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon. Language Resources and Evaluation Conference (LREC 2012)*, pages 31–36.

Efthimiou, E., Fotinea, S.-E., Hanke, T., Glauert, J., Bowden, R., Braffort, A., Collet, C., Maragos, P., and Lefebvre-Albaret, F. (2012). Sign language technologies and resources of the dicta-sign project. In *5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon. Language Resources and Evaluation Conference (LREC 2012)*, pages 37–44.

Elliott, R., Glauert, J. R. W., Jennings, V., and Kennaway, J. R. (2004). An overview of the sigml notation and sigmlsigning software system. In *Sign Language Processing Satellite Workshop of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 98–104.

Hanke, T. (2004). Hamnosys-representing sign language data in language resources and language processing contexts. In *Sign Language Processing Satellite Workshop of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 1–6.

Hanke, T. (2006). Towards a corpus-based approach to sign language dictionaries. In *Proceedings of a Workshop on the representation and processing of sign languages: lexicographic matters and didactic scenarios (LREC 2006)*, pages 70–73.

Hrúz, M., Krňoul, Z., Campr, P., and Müller, L. (2011). Towards automatic annotation of sign language dictionary corpora. In Ivan Habernal et al., editors, *Text, Speech and Dialogue*, pages 331–339, Berlin, Heidelberg. Springer Berlin Heidelberg.

Kennaway, R. (2002). Synthetic animation of deaf signing gestures. In Ipke Wachsmuth et al., editors, *Gesture and Sign Language in Human-Computer Interaction*, pages 146–157, Berlin, Heidelberg. Springer Berlin Heidelberg.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Yoshua Bengio et al., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Matthes, S., Hanke, T., Regen, A., Storz, J., Worseck, S., Efthimiou, E., Dimou, A.-L., Braffort, A., Glauert, J., and Safar, E. (2012). Dicta-sign – building a multilingual sign language corpus. In *Proceedings of 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon (LREC 2012)*, pages 117–122.

Ong, S. C. W. and Ranganath, S. (2005). Automatic sign language analysis: A survey and the future beyond lexical meaning. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 873–891.

Östling, R., Börstell, C., and Courtaux, S. (2018). Visual iconicity across sign languages: Large-scale automated video analysis of iconic articulators and locations. *Frontiers in psychology*.

Prillwitz, S., Leven, R., Zienert, H., Hanke, T., and Henning, J. (1989). *HamNoSys version 2.0. Hamburg notation system for sign languages—an introductory guide*. Signum-Verlag.

Prillwitz, S., Hanke, T., König, S., Konrad, R., Langer, G., and Schwarz, A. (2008). DGS corpus project-development of a corpus based electronic dictionary German Sign Language / German. In *3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora (LREC 2008)*, pages 159–164.

Simon, T., Joo, H., Matthews, I. A., and Sheikh, Y. (2017). Hand keypoint detection in single images using multiview bootstrapping. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4645–4653. IEEE Computer Society.

Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 217–223
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

Cross-Lingual Keyword Search for Sign Language

Nazif Can Tamer, Murat Saraçlar
Boğaziçi University
Department of Electrical and Electronics Engineering
{can.tamer, murat.saraclar}@boun.edu.tr

Abstract

Sign language research most often relies on exhaustively annotated and segmented data, which is scarce even for the most studied sign languages. However, parallel corpora consisting of sign language interpreting are rarely explored. By utilizing such data for the task of keyword search, this work aims to enable information retrieval from sign language with the queries from the translated written language. With the written language translations as labels, we train a weakly supervised keyword search model for sign language and further improve the retrieval performance with two context modeling strategies. In our experiments, we compare the gloss retrieval and cross language retrieval performance on RWTH-PHOENIX-Weather 2014T dataset.

Keywords: sign language recognition, cross language retrieval, context modeling, weakly supervised learning, attention

1. Introduction

Most of the existing data in sign language comes from the public media, where one finds news, shows, TV series, and movies interpreted for the Deaf in sign language. Although the amount of data available in this format is great in scale, these parallel corpora are often considered too noisy and unreliable for sign language research. The grammar and word ordering of the written/spoken language and the corresponding sign language interpreting do not match one to one, and thus, translations in the written language cannot be directly used to train current automatic recognition systems that require at least ordered glosses as the label.

Since the effective utilization of these parallel corpora would greatly increase the overall number of data available for sign language studies, researchers actively try to convert this weakly supervised and noisy data into a more convenient format for research. Pfister et al. (2013) and Kelly et al. (2011) use the weak supervision coming from the translations to automatically extract isolated signs with multiple instance learning (MIL) based strategies, and train models with the segmented data. In a different direction, Camgoz et al. (2018) dropped segmentation out of the equation and applied state-of-the-art neural machine translation approaches to directly translate sign language videos to the written language in an end-to-end manner. In this work, by utilizing a similar strategy, we train an end-to-end keyword search model by searching the sign language sentence for words coming from the translations in written language.

Although keyword search is a new application for sign languages, it is a well-studied problem for spoken languages. The most common strategy is to use lattices generated by automatic speech recognition (Saraclar and Sproat, 2004). More recently, end-to-end keyword search strategies also started to appear (Audhkhasi et al., 2017). We previously used end-to-end methods for gloss search from sign language videos in (Tamer and Saraçlar, 2020), and this work is an extension of that.

The main contribution of this work on top of our previous model is the introduction of the context modeling for cross-lingual keyword search. Rescoring keyword search predictions with the predictions for other keywords (Karakos et al., 2013) and the predictions of the same keyword at another close time instant (Richards et al., 2014) is a known strategy in spoken keyword search. In this work, we apply this rescoring strategy to our model's cross-lingual keyword search and show that model's own predictions for other keywords can be used to boost keyword search performance.

The rest of this paper is organized as follows. In Section 2, the previously-introduced end-to-end keyword search network is summarized briefly. In Section 3, the modifications made specifically for cross-lingual search is explained. In Section 4, the dataset and evaluation metrics are given. Lastly, in Section 5, in addition to giving our results for keyword search and comparing them to gloss search, we further discuss how this weakly supervised training strategy helps automatic segmentation of parallel corpora between written language and sign language interpreting.

2. Weakly Supervised Keyword Search for Sign Language

The model structure is summarized in Figure 1. After the video is converted into a sequence of skeleton joints, the rest of the keyword search model is trained end-to-end by searching for text or gloss queries in the sign language sentence. In short, the aim of this training strategy is to represent both a query and the relevant part of the sign language sentence by a similar vector in a mutual latent space. This is done by the joint training of spatio-temporal graph convolutional network (ST-GCN) encoder, word embedding, and the attention based selection mechanism.

2.1. ST-GCN Encoding of the Sign Language Sentence

Spatial Temporal Graph Convolutional Networks (Yan et al., 2018) first introduced for the skeleton-based action recognition is used for the encoding of the skeleton sequence. In this model, a graph connecting neighboring skeleton joints and the same joints across frames (see the

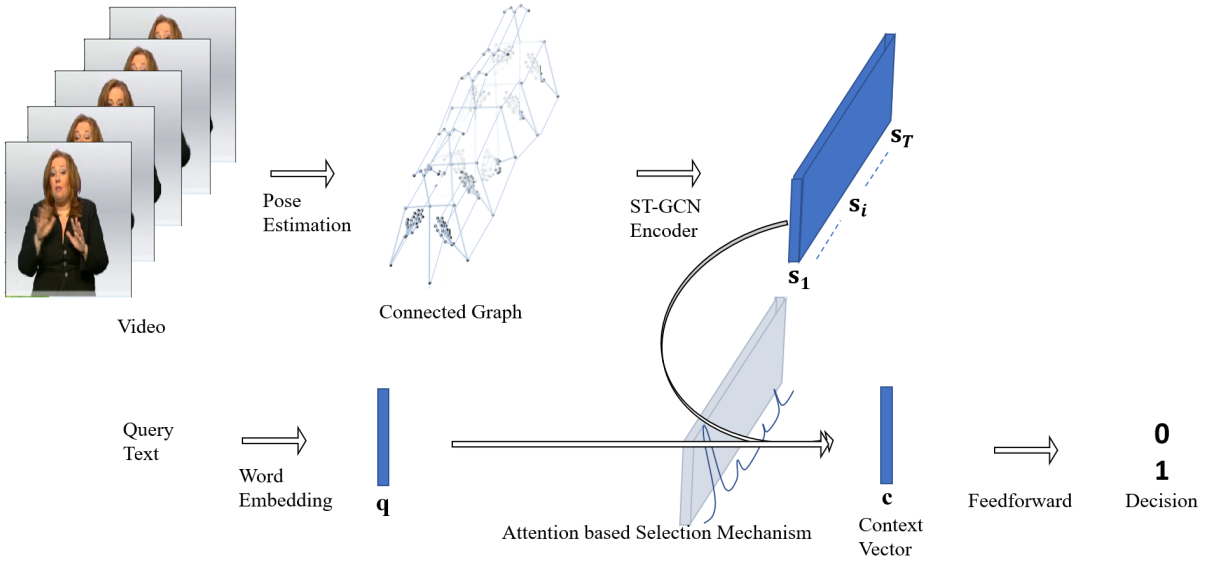


Figure 1: After pose estimation with OpenPose, the rest of the framework is trained end-to-end.

connected graph in Figure 1) is formed and 12 layers of graph convolution operations take place on top of this connected graph as described in our previous work (Tamer and Saraçlar, 2020).

2.2. Query Embedding and the Attention-based Selection Mechanism

Query embeddings, vectors representing each query in our vocabulary, are learned through attention based selection mechanism. Let $\mathbf{q}$ represent the word embedding for the query and $\mathbf{s}_i$ the ith member of the encoded sign language sentence; the similarity score between $\mathbf{q}$ and $\mathbf{s}_i$ is obtained for all $i \in (1, T)$ through the scoring function

$$score(\mathbf{q}, \mathbf{s}_i) = \beta \left[\frac{\mathbf{q} \cdot \mathbf{s}_i}{\|\mathbf{q}\| \cdot \|\mathbf{s}_i\|} \right]^2 + \theta \qquad (1)$$

with β and θ learnable parameters. From the similarity scores, a single context vector $\mathbf{c}$ is obtained

$$\mathbf{c} = \sum_i \left[\frac{\exp\left(score(\mathbf{q}, \mathbf{s}_i)\right)}{\sum_{i'} \exp(score(\mathbf{q}, \mathbf{s}_{i'}))} \right] \cdot \mathbf{s}_i \qquad (2)$$

and a simple fully connected layer decides on whether the query $\mathbf{q}$ is found in the entire $\mathbf{s}_{1:T}$ sequence.

2.3. Training Strategy

The stability of the training is ensured by searching for all the queries in our vocabulary in the same sign language sentence. With the skeleton sequence and all the queries in our vocabulary at the input, the network is trained to minimize the binary cross entropy loss between its predictions for each query and the labels obtained by simply giving 1 if the query is in the translation and 0 if it is not.

3. Query-Specific Context Modeling for Cross-Lingual Keyword Search

The motivation behind query-specific context modeling is that, when doing cross-lingual retrieval, words in the spoken language do not match one to one with their sign language glosses. Furthermore, for some less frequent words, the predictions we get by training with only small amount of data are most often unreliable. To remedy these problems, we define two prediction rescoring strategies that use model's own predictions for other queries within the vocabulary V. For a single sign language sentence, let $\vec{l}$ represent the $|V|$ dimensional correct labels, s.t. we have one label $\in \{0, 1\}$ for each query, and $\vec{p}$ represent the $|V|$ dimensional vector comprised of the trained model's predictions. Our aim is to come up with a new predictions vector $\vec{\tilde{p}}$ that is better than the original $\vec{p}$. We do this by two different strategies: (i) a statistical context model based on bag-of-words TF-IDF vectors, and (ii) a machine-learning based multi-layer perceptron (MLP) context model.

3.1. Statistical Context Modeling with TF-IDF Vectorization

Term Frequency Inverse Document Frequency (Ramos and others, 2003) vectorization is a well known strategy for language modeling. While calculating a weight for a query inside a document, this algorithm gives high weights to queries that are seen multiple times in this document (high term frequency), and low weights to ones that are seen in many other documents (inverse document frequency). Thus, by finding the similarity between each keyword in the vocabulary V and the document, we obtain a $|V|$ dimensional vectoral representation of the document. Let $\vec{d}_i$ be the l_1 normalized TF-IDF vector for the ith document in our training set, the document context model $\vec{d}_q$ for query q is found by averaging over all the documents in our training set that contain this specific query:

$$\vec{d}_q = \text{avg}(\vec{d}_i : \text{tfidf}(q, \vec{d}_i) > 0) \qquad (3)$$

Then, by looking at the cosine similarity between this query-specific document context vectors $\vec{d}_q$ and our

model's prediction vector $\vec{p}$, we obtain the new scores. The new prediction score for the ith query in the vocabulary $\tilde{p}(q_i)$ is formulated as

$$\tilde{p}(q_i) = 1 - \frac{\vec{d_q} \cdot \vec{p}}{|\vec{d_q}||\vec{p}|} \tag{4}$$

and the new prediction vector $\vec{\tilde{p}}$ is simply the new prediction values for all the queries in our vocabulary V.

$$\vec{\tilde{p}} = \left[\tilde{p}(q_1), ..., \tilde{p}(q_i), ..., \tilde{p}(q_{|V|})\right]^\intercal \tag{5}$$

3.1.1. Fusion Strategy

The statistical context modeling for the query, by itself, does not give better results than the model's own predictions. However, when combined with the original predictions through a hyperparameter, it boosts the prediction scores. With $\vec{p}$ being the model's original predictions and $\vec{\tilde{p}}$ the predictions obtained through query context modeling, the final predictions are obtained by combining the two predictions using a hyper-parameter γ:

$$\log \vec{\tilde{p}}' = \gamma \cdot \log \vec{p} + (1 - \gamma) \cdot \log \vec{\tilde{p}} \tag{6}$$

In our experiments, we tuned the hyperparameter γ to maximize the mean average precision (mAP) score in the development set. For different graph layout options, results given at Table 2 are with the γ values of 0.40 for the upper body only layout, 0.54 for the upper body and dominant hand combined, and 0.58 for all upper body, the dominant hand and the passive hand combined.

3.2. Multilayer Perceptron Based Context Model

For a sign language sentence in our training set, we trained a simple multi-layer perceptron with $|V|$ dimensional predictions vector $\vec{p}$ as the inputs and labels vector $\vec{l}$ as the target. The network is comprised of two hidden layers of size 256 with ReLU activations and a dropout probabilty of 20%. We finished the training with early stopping when the loss in the development set was not reducing any further.

4. Experimental Setup

4.1. Dataset

We used RWTH-PHOENIX-Weather-2014T dataset (Camgoz et al., 2018) to conduct our experiments. Recorded in 25 fps videos, the dataset includes weather forecasts in sign language, their sentence-level gloss transcriptions (without temporal alignments), and the translations into the German language. The main reason we used this dataset for our experiments is that, by including both gloss transcriptions in German sign language and corresponding translations in German, it offers a natural medium for comparing cross-lingual keyword search with gloss search.

The dataset is partitioned into 9.2 hours of training, 37 minutes of development and 43 minutes of test data. In order to use this dataset in keyword search task, we segmented the transcriptions and translations into constituent words and used them as our queries. In the gloss search, the vocabulary consists of 1085 glosses that are seen at least once in the training set and 398 of these are also seen at least once in the test dataset. Thus, we report our results from this shared vocabulary of $\mathbf{398}$ queries. Similarly, for the cross-lingual search, we have 2887 words in the training set and 942 of these are also shared in the test set and we report our cross-lingual results on this shared vocabulary of $\mathbf{942}$ queries.

To clarify the training procedure with an example, let us consider the sequence in Figure 6: When training a gloss search model, a 40-frame long sign language sentence is labeled with $-1-$ for 3 glosses: "nordost", "bleiben" and "trocken", and $-0-$ for the remaining 1082 glosses. When training a cross-lingual keyword search model, the same sequence is labeled with $-1-$ for 6 words: "im", "nordosten", "bleibt", "es", "meist", "trocken", and $-0-$ for the remaining 2881 words. A cross-lingual kws model cannot see the glosses and vice versa; gloss and cross-lingual search models are completely independent.

4.1.1. Skeleton Extraction from Video Frames

2D pose estimates of upper body, right and left hand are extracted through part affinity fields based OpenPose framework (Cao et al., 2017). In figure 2, you can see an example subsequence from a sign language sentence with OpenPose pose estimates projected on top. Since the frames are blurry and low resolution, the pose estimation process cannot always result in good (x, y) coordinate estimates for each joint. To remedy this, we also used the related confidence scores as the third dimension to feed into the graph convolutional encoder.

4.2. Evaluation metrics

For a query q, precision recall values at an operating point are defined as

$$\text{Precision} = \frac{|\{\text{Retrieved}\} \cap \{\text{Relevant}\}|}{|\{\text{Retrieved}|\}}$$
$$\text{Recall} = \frac{|\{\text{Retrieved}\} \cap \{\text{Relevant}\}|}{|\{\text{Relevant}\}|}$$

and precision-recall curve obtained at different operating points (e.g. by changing the threshold) is one of the most valuable metrics in evaluating the performance of information retrieval systems.

4.2.1. Term-averaged Precision-Recall Curve and the F1 Score

When precision and recall values associated with a threshold θ is averaged over different queries q, term-averaged precision-recall values are obtained for that threshold:

$$\text{Precision}(\theta) = \frac{1}{|Q|} \sum_{q \in Q} \text{Precision}(q, \theta)$$
$$\text{Recall}(\theta) = \frac{1}{|Q|} \sum_{q \in Q} \text{Recall}(q, \theta)$$

Thus, by sweeping through different θ thresholds, we obtain the term-averaged precision-recall curve that summarize the performance of the keyword search system. We also report the maximum of F1 scores summarizing the curve:

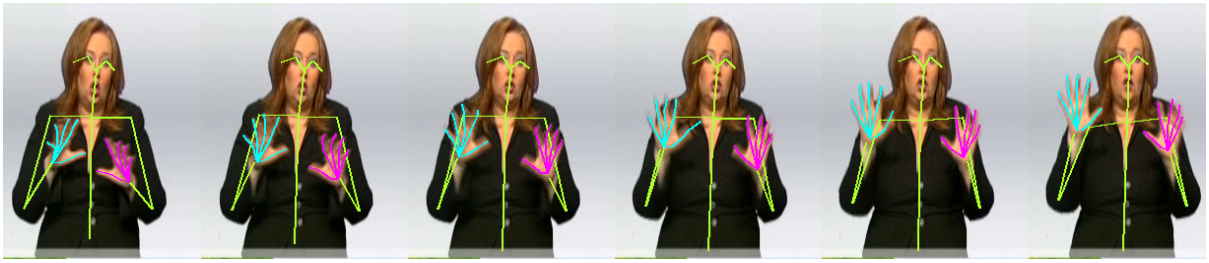


Figure 2: An example subsequence from the dataset. The extracted poses for upper body, the dominant hand, and the passive hand are shown on top of images in yellow, cyan, and magenta respectively. The poses constructed by OpenPose framework are highly representative even though the original input images are low resolution and blurry.

$$\text{maxF1} = \max_{\theta} \frac{2 \cdot \text{Precision}(\theta) \cdot \text{Recall}(\theta)}{\text{Precision}(\theta) + \text{Recall}(\theta)}$$

4.2.2. Mean Average Precision (mAP)

Similarly, in object and action recognition, one of the most used metrics is mean average precision. It roughly corresponds to the area under precision-recall curves belonging to different queries q averaged over queries.

$$\begin{aligned}\text{mAP} &= \frac{1}{|Q|} \sum_{q \in Q} \text{AveragePrecision}(q) \\ &= \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{|N|} \sum_{n=1}^{|N|} \text{Precision@}n(q)\end{aligned}$$

($|N|$: the number of relevant documents for query q). It is common to report mAP scores at different Intersection over Union (IoU) thresholds. However, since we did not have any labels for temporal alignments and segmentation, we simply report mAP scores with IoU=0.

5. Results and Discussion

In this section, we present our gloss and cross-lingual keyword search results obtained with different encoder graph structures and different context modeling strategies. We also compare cross-lingual KWS results obtained with a translation approach and visualize the temporal localization capabilities of our model.

5.1. Effects of Graph Layout: Upper Body, the Dominant and the Passive Hand

The upper body, the dominant hand, and the passive hand poses are all important components in understanding sign language. To identify the effects of different components in the performance of keyword search for sign language, we trained 3 gloss search and 3 cross-lingual keyword search models with the features in Figure 3.

From the results summarized in Table 1 and the precision-recall curve in Figure 4, we see that the upper body alone contains much of the information by itself. Introducing the dominant hand also significantly improves the results for both gloss and cross-lingual search models. However, we see that there is not much gain with the introduction of passive hand. Although the layout including the passive hand performs the best in cross-lingual search, it reduces both the mAP and maxF1 scores in the gloss search compared to the dominant hand + upper body layout option.

Since the OpenPose hand model has 21 joints, including the passive hand in the graph layout increases the number of graph nodes from 34 to 55 and demands more computational resources for graph convolution operations. Thus, we conclude that the costs of including the passive hand in the graph layout may outweigh the benefits.

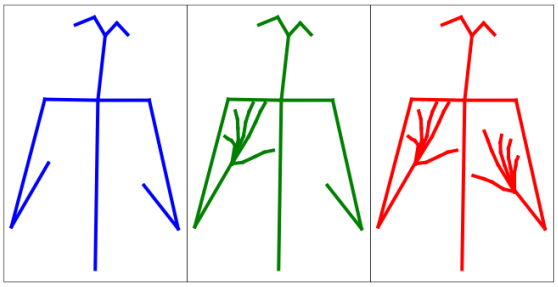


Figure 3: From left to right, the three graph layout options used in the experiments are upper body (13 joints), upper body with the dominant hand (34 joints), and upper body with both hands (55 joints), respectively.

	Gloss		Cross-Lingual	
	mAP (%)	maxF1	mAP(%)	maxF1
Upper Body (UB)	24.29	26.40	12.49	15.18
UB + Dom. Hand	**29.91**	**33.53**	13.18	15.62
UB + Both Hands	29.22	32.80	**14.56**	**16.15**

Table 1: Gloss and cross-lingual KWS results using two metrics (the higher the better, best scores for each task are in bold). Cross-lingual results are reported after MLP context model applied.

5.2. Effect of Different Context Modeling Strategies

Results obtained with different context modeling strategies are summarized in Table 2. Firstly, we can say that sta-

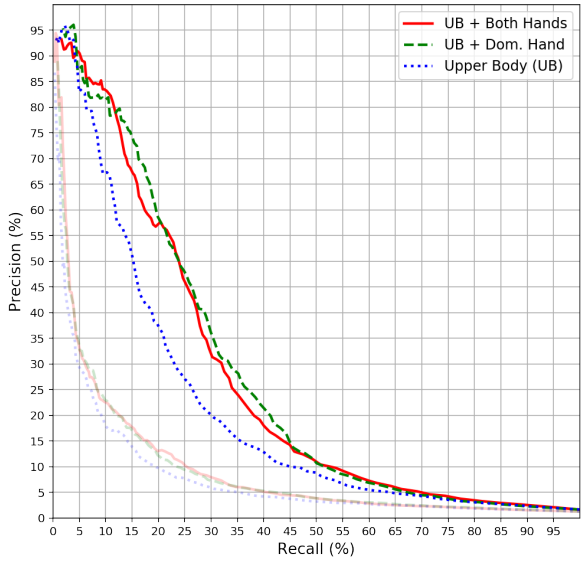


Figure 4: Precision-recall curves for the gloss search models with different layout options. The cross-language search results are shown in transparent for comparison.

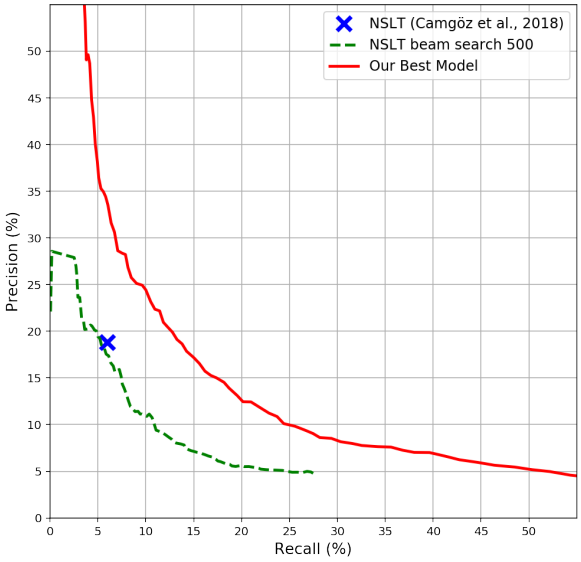


Figure 5: Our best cross-lingual KWS model (trained with UB + Both Hands layout option and MLP context model) compared to searching from Neural Machine Translation outputs (the higher the better).

tistical context modeling improves both metrics for all the layout options, and the gains are significant for the Upper Body + Both Hands layout. Secondly, we see that the MLP based context modeling did not improve the results for UB + Dominant Hand layout. Since we stopped the training of the MLP based context model when the development loss is not reducing any further, the results in the test dataset are not necessarily better. However,we obtained our best overall mAP score with an MLP based context model (with a significant increase from 13.01% to 14.56%).

5.3. Comparison to KWS from Neural Machine Translation Outputs

In spoken keyword search, a well-known strategy is to use the transcriptions obtained through automatic speech recognition (ASR). In a similar approach, we used translations obtained from Neural Sign Language Translation (Camgoz et al., 2018) model as our baseline. From the translations we get by using the same hyper-parameters in their paper, we obtain the single operation point denoted as NSLT in Figure 5.

In spoken keyword search, another strategy is to search for the keyword in lattices generated from ASR outputs (Saraclar and Sproat, 2004). Similarly, we plot precision-recall curve related to this NSLT model by applying beam search with beam size of 500 and finding the expected counts for each word along the beams. With the two as our baselines in Figure 5, we conclude that our cross-lingual KWS model is better than searching for keywords in translation outputs.

5.4. Temporal Localization as a By-Product of Weakly Supervised Training

When we have sequence-level, ordered gloss transcriptions of sign language data, HMM-based models can iteratively

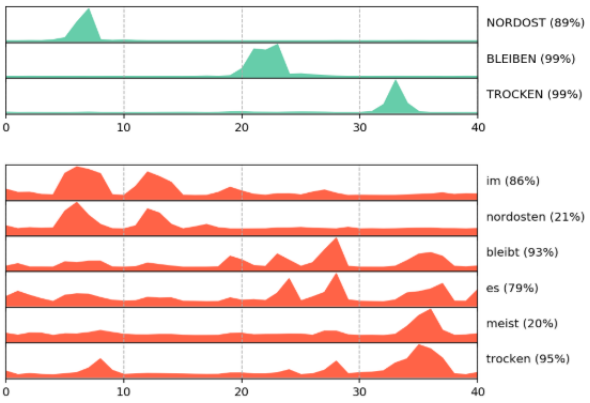


Figure 6: Temporal localizations for the sequence with gloss annotation “nordost bleiben trocken” and translation “im nordosten bleibt es meist trocken”. The prediction confidences are denoted in parentheses.

align each frame to a gloss hidden state and thus do the temporal segmentation as exemplified in (Koller et al., 2017). However, since these HMM models rely on the strictness of the order of a gloss sequence, this alignment procedure cannot work with the noisy and weak supervision of translations. In this section, we show that our model’s attention based selection mechanism can loosely localize some keywords independent of label type. For sign language sentences of varying length, we show the temporal keyword localization capabilities of our models that are trained with either gloss-sequences or translations as the labels.

In Figures 6, 7, and 8, we see model predictions (shown

	Without Context Model		Statistical C.M.		MLP-based C. M.	
	mAP (%)	maxF1	mAP (%)	maxF1	mAP (%)	maxF1
Upper Body (UB)	11.62	14.66	11.94	14.90	**12.49**	**15.18**
UB + Dom. Hand	13.66	16.10	**13.78**	**16.28**	13.18	15.62
UB + Both Hands	13.01	16.40	13.80	**16.69**	**14.56**	16.15

Table 2: Effect of context model on cross-lingual KWS. Best mAP and maxF1 scores for each layout are in bold, and overall best scores are underlined.

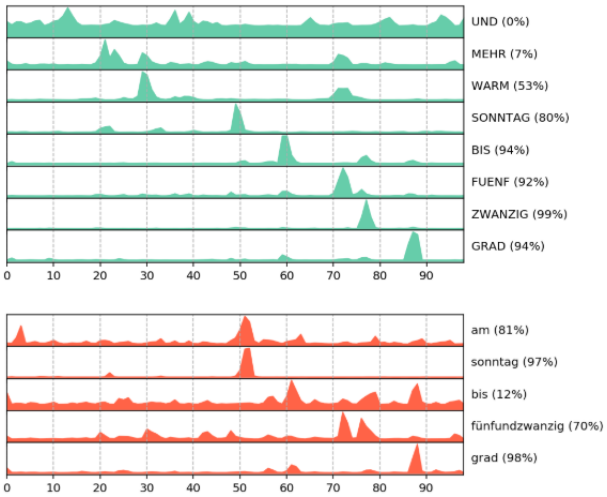


Figure 7: Temporal localizations for the sequence with gloss annotation "und mehr warm sonntag bis fuenf zwanzig grad" and translation "am sonntag bis fünfundzwanzig grad". The prediction confidences are denoted in parentheses.

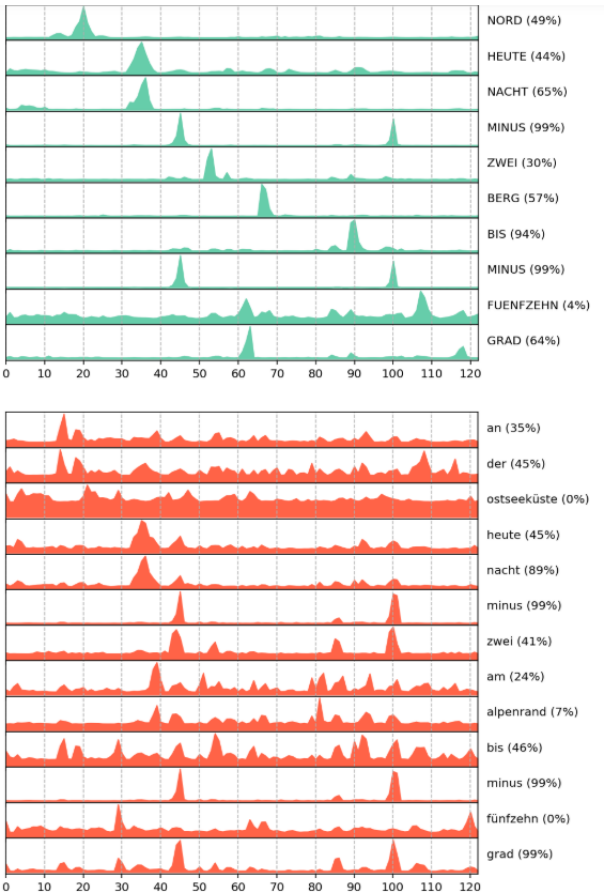


Figure 8: Temporal localizations for the sequence with gloss annotation "nord heute nacht minus zwei berg bis minus fuenfzehn grad" and translation ""an der ostseeküste heute nacht minus zwei am alpenrand bis minus fünfzehn grad". The prediction confidences are in parentheses.

with percentages next to the labels) and related temporal localizations (denoted by the most peaky regions) for both gloss and cross-lingual search. For the most of our data, we see that gloss search models are better in localization capacity and the order of peaky regions usually follows the gloss order correctly. We also see that peaky regions are more visible when the prediction confidences are higher. For the the cross-lingual search, we see that localization is possible for some words that are matching one-to-one with gloss transcriptions (such as "grad" in Figures 7 and 8, "sonntag" and "fünfundzwanzig" (with two peaks at both "fuenf" and "zwanzig") in Figure 7, "nacht" in Figure 8 etc.), but not so much for the conjugated verbs like "bleibt" in Figure 6, or words without a unique gloss such as "alpenrand" and "ostseeküste" in Figure 8. We believe that cross-lingual KWS is at least beneficial for finding the most salient temporal regions that might be related to any gloss.

6. Conclusion

In this paper, we employed a weakly-supervised, end-to-end training strategy for cross-lingual keyword search for sign language and showed that cross-lingual training is a viable option when we do not have the gloss labels. We introduced two context modeling strategies and further improved the cross-lingual keyword prediction results. We compared the retrieval performance and temporal localization capabilities of gloss and cross-lingual search under three different layout options. The most important contribution of this paper is the introduction of a cross-lingual KWS method that can theoretically utilize the widely available sign language interpretations in public media. In the future, we aim to apply the same strategy to bigger datasets.

7. Acknowledgements

This study was supported in part by the Scientific and Technological Research Council of Turkey (TUBITAK) under Project 117E059.

8. Bibliographical References

Audhkhasi, K., Rosenberg, A., Sethy, A., Ramabhadran, B., and Kingsbury, B. (2017). End-to-end asr-free keyword search from speech. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1351–1359.

Camgoz, C., Hadfield, S., Koller, O., Ney, H., and Bowden, R. (2018). Neural sign language translation. In *Proc. CVPR*, pages 7784–7793.

Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*.

Karakos, D., Schwartz, R., Tsakalidis, S., Zhang, L., Ranjan, S., Ng, T., Hsiao, R., Saikumar, G., Bulyko, I., Nguyen, L., Makhoul, J., Grézl, F., Hannemann, M., Karafiát, M., Szoke, I., Veselý, K., Lamel, L., and Le, V. B. (2013). Score normalization and system combination for improved keyword spotting. pages 210–215, 12.

Kelly, D., McDonald, J., and Markham, C. (2011). Weakly supervised training of a sign language recognition system using multiple instance learning density matrices. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 41(2):526 –541, april.

Koller, O., Zargaran, S., and Ney, H. (2017). Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent cnn-hmms. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July.

Pfister, T., Charles, J., and Zisserman, A. (2013). Large-scale learning of sign language by watching tv (using co-occurrences). In *BMVC*.

Ramos, J. et al. (2003). Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 133–142. Piscataway, NJ.

Richards, J., Ma, M., and Rosenberg, A. (2014). Using word burst analysis to rescore keyword search candidates on low-resource languages. pages 7824–7828, 05.

Saraclar, M. and Sproat, R. (2004). Lattice-based search for spoken utterance retrieval. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 129–136.

Tamer, N. C. and Saraçlar, M. (2020). Keyword search for sign language. In *Proc. ICASSP*.

Yan, S., Xiong, Y., and Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence*.

Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 224–230
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

One Side of the Coin: Development of an ASL-English Parallel Corpus by Leveraging SRT Files

Rafael O. Treviño[1], Julie A. Hochgesang[2], Emily P. Shaw[1], Nic Willow[2]
[1]Department of Interpretation and Translation, [2]Department of Linguistics
Gallaudet University
800 Florida Ave NE, Washington, DC 20002, USA
{rafael.trevino, julie.hochgesang, emily.shaw, nic.willow}@gallaudet.edu

Abstract
We report on a method used to develop a parallel corpus of English and American Sign Language (ASL). The effort is part of the Gallaudet University Documentation of ASL (GUDA) project, which is currently coordinated by an interdisciplinary team from the Department of Linguistics and the Department of Interpretation and Translation at Gallaudet University. Creation of the parallel corpus makes use of the available SRT (SubRip Subtitle) files of ASL videos interpreted into or from English, or captioned into English. The corpus allows for one-way searches based on the English translation or interpretation, which is useful for translators, interpreters, and those conducting comparative analyses. We conclude with a discussion of important considerations for this method of constructing a parallel corpus, as well as next steps that will help to refine the development and utility of this type of corpus.

Keywords: corpus, parallel corpus, translation, interpreting, SRT

1. Introduction

The method of constructing a corpus presented in this paper addresses two issues. The first is in constructing a corpus. Even with written text, which is in a machine-readable format, constructing a corpus can be laborious and time-consuming. Sign language corpora are all the more difficult due to the need to adopt conventions for converting the data from video into a machine-readable format. In some sign language corpora, data is sometimes freely translated into a written language as a way to provide provisional access to the signed language. This translation task, however, still takes time that an annotator could use to work on annotation.

The second issue relates to the use of corpora by those from other disciplines. In looking toward the future in their text on corpus linguistics, McEnery and Hardie (2012) offer ideas on "the potential for corpus methods to extend beyond the field of linguistics into other areas of the humanities, sciences and social sciences" (p. 225). In our case, we exhibit an area for collaboration between sign language corpus linguistics and the field of sign language translation and interpreting. For example, when faced with challenges in how to translate a specialized term or phrase from English into ASL, students have had to rely on personal observations of deaf people and interpreters to build their vocabulary and repertoire of interpretation choices. However, their observations are fleeting (i.e., cannot be accessed later for review) and limited in number and kind.

Thus, the parallel corpus described in this paper presents a possible solution to the issues raised here. In the first case, our proposed method can be used to leverage SRT files to save annotators time and allow linguists (or others) to establish a provisional corpus expeditiously. In the second case, the resulting parallel corpus, while it cannot be exploited in its initial state by linguists, the parallel corpus can be used by others who can profit from the ability to investigate the signed language through another, written language. In other words, one linguist's provisional corpus is another field's treasure.

In addition, we must point out the two cases can work in tandem: a provisional corpus can be used by people other than linguists while the signed language side of the corpus is being annotated.

In any event, the parallel corpus described in this paper provides the means to conduct powerful analyses of larger interpreted datasets. We suspect, moreover, it may have applications beyond the suggestions and ideas presented herein.

2. Background

2.1 Gallaudet University Documentation of ASL (GUDA) Project

Gallaudet is an ASL-English bilingual university. The campus community consists of Deaf, DeafBlind, hard of hearing, and hearing people, all of whom have varying degrees of fluency in ASL (visual and tactile varieties), written and spoken English, not to mention other written, spoken, and signed languages. Because of its bilingual mission (Gallaudet University, 2007), the university commits to providing video content of lectures, announcements, and other communications in both ASL and English (written or spoken, or both).

The creation of the parallel corpus emerged from work on the GUDA project. The project aims to digitally organize the ASL video collections on campus so they may be accessed by scholars and the public (see Hochgesang, Willow, Treviño, and Shaw, 2019, for a more complete description of the project). Notably, the research team behind the project is currently composed of faculty and graduate students from both the Department of Linguistics and the Department of Interpretation and Translation. It is partly the intersecting interests of these two disciplines that helped uncover the benefits of combining the needs of translators and interpreters with the technology for building sign language corpora.

2.2 Definitions

Later in the paper, we present some important considerations regarding terminology in the face of multi-modal parallel corpora. For the moment, however, it may be useful to the reader for us to review a few preliminary terms.

In this paper, *translation* refers to the act of rendering ASL in a video into written English after the recorded event has occurred. The written English may appear either as subtitles or as a tier in ELAN, or both. This activity can be carried out by an annotator, a professional, or, in the case of some videos, an unknown person.

By *interpreting* or *interpretation*, we refer to the act of rendering either ASL into English or English into ASL, most often in the simultaneous mode.

We typically use *transcription* to refer to a representation of spoken English in written form. Transcription can be used to represent spoken English, either as the source message or as the interpretation of an ASL source message. When a transcription is provided at the time of the recorded event, we refer to this as *real-time transcription*. When a transcription is produced after the event has already taken place, we refer to it simply as *transcription*, or *offline transcription*.

2.3 Sign Language Corpora as Parallel Corpora

Baker, Hardie, and McEnery (2006) define a *parallel corpus* as "a set of texts and their translations" (p. 126). They note parallel corpora are often used to compare terms and grammatical structures between languages, to look at the features of translations, and to assist with machine translation.

Sign language corpora often include translation into a written language as one of the steps in converting the data into a machine-readable format. Meurant, Cleve, and Crasborn (2016) observe that this work to translate the signed language into a written language effectively converts sign language corpora into bilingual (a.k.a., "parallel") corpora.

Indeed, Meurant, Cleve, and Crasborn (2016) also emphasize that the bilingual nature of signed language corpora has yet to be fully exploited for purposes such as the ones noted by Baker, Hardie, and McEnery (2006). They draw upon the Corpus LSFB (Meurant, 2015) and the Corpus NGT (Crasborn, Zwitserlood, & Ros, 2008) to describe how linguists, interpreters, translators, teachers, and language learners can all use parallel corpora of signed languages. At the time of the Meurant, Cleve, and Crasborn (2016) paper, the Corpus LSFB had 2.5 hours (2,400 sentences) of LSFB with translations into French, and the Corpus NGT had 15 hours (15,000 sentences) of NGT with translations into Dutch.

2.4 Corpora and Sign Language Interpreting

Parallel sign language corpora are not just for looking at questions pertaining to sign language interpreting, but the two do seem to go hand-in-hand. In Translation and Interpreting Studies (TIS), the benefits of corpora have been recognized since at least the 1990s, especially with regard to investigating theoretical issues and the process of interpreting (Baker, 1993; Shelsinger, 1998). With regard to sign language interpreting, the benefits of using corpora have also been recognized, mostly in the realm of training interpreters. Early on, for instance, Heßmann and Vaupel (2008) argued for the need to implement the use of sign language corpora in interpreter education and outlined some of the challenges in doing so. One of the challenges was taking into consideration spoken language, even though it may "seem odd to include vocal language texts in a sign language corpus" (p. 75). However, when creating a parallel corpus for comparative purposes, it seems handling the data of spoken language cannot be ignored. To this end, Heßmann and Vaupel (2008) also identified challenges related to the classification of data and the use of metadata in parallel corpora, which we also address in our paper in section 4.2.

As an invited speaker from outside linguistics at a workshop given for the Sign Linguistics Corpora Network in Berlin, Nancy Frishberg (2010) mentioned a few of the possibilities corpora hold for interpreter education. She posited corpora could be used "within mode" for non-native users of sign language to improve their linguistic ability and "across modes" for comparative analyses. The importance of corpora as a vehicle is that they provide the advantage of annotations, which enable the data to be searched (e.g., "I want to work on conversations, especially those with head-tilt" [slide 33]). Moreover, the benefit could be mutual in that the learners themselves could also provide input to the annotations (i.e., to crowdsource the annotation effort).

Since Heßmann and Vaupel (2008) and Frishberg (2010), some progress has been made in developing corpora for sign language interpreting, but there are only a few references in the literature. Wehrmeyer (2019) provides an in-depth account of her work in constructing the South African Sign Language Interpreting Corpus (SASLIC), a parallel corpus of English–South African Sign Language (SASL) based on interpreted news bulletins. The English source text (ST) of the corpus was created using re-speaking software, and the SASL target text (TT) was annotated using a novel convention. Wehrmeyer (2019) notes that, "a typical half-hour of ST could be transcribed in a day, whereas the TT transcription for that selection took at least 160 hours" (p. 73). In her paper, she reports a total of approximately 3.3 hours (200 minutes) of source text that had been transcribed. She concludes by observing sign language parallel corpora such as hers could be used to investigate "referencing techniques, non-manual features, discourse devices and interpreting strategies" (p. 81).

In another parallel corpus, Roush (2016) compiled the translations produced by deaf translators from English into ASL of famous speeches in U.S. history. The corpus, known as the American Freedom Speeches (AFS) Translation Corpus, consists of 29 minutes of video. The English source text was fully transcribed and the ASL target text was fully annotated. The purpose of the AFS corpus was to explore its pedagogical utility in teaching sign language interpreters. Roush (2016) notes one of the advantages of the AFS corpus is that it shows learners how native users of ASL expressed particularly problematic constructions in the English source into ASL.

The foregoing are two examples of corpora developed for broad purposes, with the former having a pedagogical slant. Due to technological and other constraints, neither is publicly accessible on a website. It is also quite possible, if not certain, other small-scale parallel corpora exist, even if they have only been compiled on an *ad hoc* basis to answer specific research questions. A case in point is the well-known study conducted by Cokely (1986). In his study, Cokely (1986) recorded, transcribed, and annotated a total of approximately 32 minutes of an interpretation conducted from English into ASL. From this data, Cokely (1986) identified there was a negative correlation between lag time and the number of errors committed by the interpreter. Like the other corpora mentioned above, the data used by Cokely (1986) is also not publicly accessible.

2.5 Summary of Issues

Several issues related to parallel corpora for sign language interpreting have been raised in this brief section. In comparison to sign language corpora in general, parallel corpora are fewer in number and smaller in size, lack a common framework for classification and metadata, and their utility is still open to possibilities. Though often geared toward pedagogical purposes, parallel corpora can

also be used to research theoretical issues and the process of interpreting. In sum, if sign language corpora are now starting to come of age, then sign language interpreting parallel corpora are the younger sibling toddling behind, tugging at the sleeves.

3. The ASL-English Parallel Corpus

As part of the GUDA project, our research team has access to an archive of over 2,000 videos held by Gallaudet University's library service and other departments on campus. Work is underway to annotate the ASL that appears in the videos in order to analyze the data from a linguistic perspective (see Hochgesang, Crasborn, and Lillo-Martin, 2018, for a review of our ASL annotation principles). One of the steps in annotating the videos includes the creation of a "free translation" tier in ELAN (Version 5.8) into English of the ASL that appears in the video. During our process of cataloging the videos available for annotation, we observed that certain collections of the ASL videos had English subtitles. We attempted to automatically create the free translation tier based on the subtitles, which were readily available. We outline this attempt in this section.

3.1 Materials

Materials for the corpus consisted of videos housed by Gallaudet University's library service with an available SRT (SubRip Subtitle) file. An SRT file is effectively an English representation of the ASL that appears in the video. Thus, its presence in the video eliminates the need for annotators to provide the translations and allows them to focus on (the more typically time-consuming act of) annotation of the ASL instead.

After eliminating all of the videos that did not have an SRT file available, we identified 590 videos as potential candidates for our parallel corpus. Of those, five were duplicates, leaving us with 585 videos and their respective SRT files. In total, the video data equal 107.48 GB and the SRT files equal 24.7 MB. Using a rough calculation based on 1 hour of video for every 500 MB (0.5 GB), we can estimate the size of the corpus to be approximately 215 hours (107.48/0.5 = 214.96).

3.2 Construction

All scripts referred to in this section were written in Apple's Script Editor (Version 2.11) on a MacBook Pro (2017) running macOS Catalina (Version 10.15.3). We use ELAN (Version 5.8).

We first created EAF files for the video files using ELAN's batch-processing functionality (File > Multiple File Processing > Create Transcription Files for Multiple Media Files). We directed ELAN to the folder containing the 585 video files. We did not select a template for the new transcription files (but see section 4.2 for a discussion on the type of information we may want to include in a template in the future). We directed the location for the new transcription files to the same folder as the media files.

Using a master spreadsheet that contained all of the video file names, we assigned each video an ID. We wrote a script to automatically create a folder for each Video ID. We then wrote a script to automatically move the video files (almost all in MP4 format), the newly created EAF files, and the SRT files into their respective folder.

ELAN (Version 5.8) provides the functionality to import an SRT file and automatically create a tier, which it names "Subtitle-Tier," based on the text and timestamps contained in the SRT file. See the user manual for a review of this functionality (Hellwig et al., 2019, p. 91). However, ELAN will not run this functionality on a batch of files; it will only import an SRT file and attach it to the EAF that is currently open. Therefore, we wrote another script to (a) open each folder, (b) open the EAF file inside the folder, (c) import the associated SRT file, (d) close the EAF, and (e) open the next folder, and so on until the process was completed. See Figure 1 for the resulting file structure.

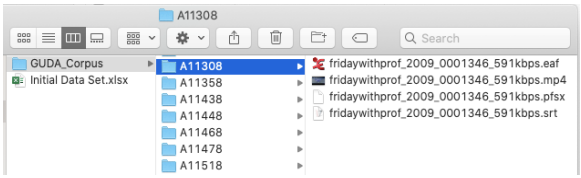


Figure 1: File structure of the ASL-English parallel corpus.

3.3 Application

At this early stage, the ASL-English parallel corpus has not been used to investigate any research questions and has only recently been introduced to interpreting and translation students. Leveraging the SRT files has served to automatically add what is effectively the free translation tier in ELAN, saving GUDA annotators a precious amount of time. Nonetheless, we would like to demonstrate one example of the utility of a parallel corpus constructed by leveraging SRT files: terminological searches.

The website HandSpeak (www.handspeak.com) is an online ASL Dictionary. A search for the English term "once" returns the ASL equivalent shown in Figure 2. The site provides two other ASL equivalents based on the usage of the word in the phrases "once a week" and "once in a while," but the initial forms are much the same as the one that appears in Figure 2.

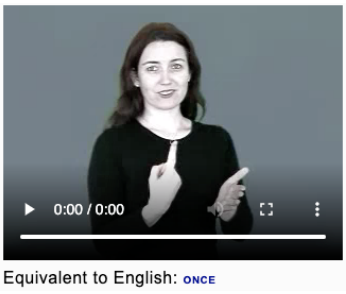


Figure 2: Word of the day (www.handspeak.com, retrieved on February 15, 2020).

By comparison, a search for the English term "once" in the ASL-English parallel corpus returns 118 occurrences, all of them in context. The results include an instance of an English interpretation — "*Once* someone transcribes video footage through [system]"—for which the ASL source text contained the sign shown in Figure 3. Note that, to ensure signs can be clearly seen and to make them accessible to the reader, we will use images from ASL Signbank (Hochgesang, Crasborn, & Lillo Martin, 2020) and include the ID gloss.

Figure 3: FINISH (ASL Signbank, 2020).

Another occurrence is from a segment of spoken English interpreted into ASL. The English segment was "because they would see an entire line of text all at *once,*" and the ASL sign corresponding to the concept of *all at once* in that context is the sign shown in Figure 4.


Figure 4: SAME-TIME (ASL Signbank, 2020).

In another result, a segment of ASL was interpreted into English as, "At school I can frequently interact as much as possible, but *once* I left school [...]." The ASL rendition did not contain a signed equivalent for "once." Rather, a shift in the signer's body posture from right to left signaled the contrast in time: before and after leaving school.

The foregoing is an abbreviated example of the potential of a parallel corpus. A few of the uses in teaching and researching sign language interpreting have already been mentioned. However, there are many more potential uses, and an exhaustive review is beyond the scope of this paper. For the time being, we ask our readers to let their imagination soar.

The creation of the ASL-English parallel corpus was a fruitful endeavor; however, it is not complete, and there are opportunities for us and others to improve upon the process. In the following section, we report on issues we encountered that we deem must be taken into consideration in the construction of a sign language parallel corpus.

4. Issues and Next Steps

4.1 One Side of the Coin

The most salient feature of this type of parallel corpus is that it is machine-readable in only one of the languages (specifically here, English). Sophisticated algorithms could probably mine the images and video that are aligned with the English text; however, most users interested in conducting a text-based search derived from the signed language, such as an ID gloss, will be disappointed. We must therefore emphasize this is a parallel corpus that can only be accessed through one of the languages. In other words, searches and analyses can only be initiated from the English representations of the ASL.

Nonetheless, those who wish to create bilingual corpora can still use the methods described herein to prepare a provisional corpus. In fact, leveraging SRT files to produce a free translation tier in ELAN does not preclude any other annotation work or linguistic analyses. All the videos used for this parallel corpus, for instance, are still in the queue for ASL annotation for the linguistic research that still needs to be done. In fact, it would be a useful time-saving strategy considering the lengthy effort of ASL annotation.

4.2 Classification

We use this section to address broad issues of classification that seem to be particular to sign language parallel corpora. We feel clear standards regarding the classification, and thus organization, of the data that feeds into parallel corpora will enable all stakeholders to take the fullest advantage of them.

4.2.1 Classification of Event Types

In section 3.1, we reported on the materials used to create this ASL-English parallel corpus, which were videos with SRT files. However, we feel more discussion is merited regarding what the SRT files represent. For the purposes of this discussion, we will use the term *online* to refer to communicative events that occurred at the time the video was recorded and *offline* to refer to events that occurred afterward.

In general, there are three types of online events that are of interest for our parallel corpus: interactions that occurred in (a) ASL only, (b) ASL interpreted into spoken English, and (c) spoken English interpreted into ASL. The English representation of the ASL that appears in the SRT can come from a number of different sources.

In Event Type A, the ASL in the video is translated offline into English, and there is time to disambiguate any utterances in the source. In the case of signed language corpora, this is a common scenario: the signed language is often translated by the annotator using the free translation tiers. In our corpus, however, there are many instances of ASL videos being translated offline into English by an unknown translator (although often carried out by a professional). The difference is not without theoretical and practical implications, as the intended audience of a translation has a significant effect on textual choices made by the translator (or annotator). For instance, a translation produced for subtitles for a public audience may synthesize information in order not to overload the screen with text. Annotators, on the other hand, do not concern themselves with how the translation will display on screen and, therefore, may approach translation differently.

Event Type B (ASL>English) represents the most complex group of the three. In straightforward cases, a verbatim transcript is produced offline of the English interpretation. However, there are many cases in which the ASL is either re-translated offline into English, or the verbatim English transcript of the interpretation is edited. In another scenario, a real-time captioner transcribes the English interpretation on-site. The online nature of real-time transcription means there is another opportunity for infelicities between the spoken English interpretation and its real-time transcription. Offline, this transcript may be used as-is to produce the subtitles or it may be edited. In sum, SRT files for Event Type B events may come from any one of three sources: 1) a re-translation of the ASL into

English, 2) a verbatim transcript of the English interpretation, or (if the service was provided) 3) a transcript of the real-time captioning based on the English interpretation, all with the possible intervention of editing.

Unlike Event Types A and B, in which the SRT file represents English as a target language, the SRT files for Event Type C (English>ASL) videos represent English as the source language. For these events, the ASL interpretation is not typically back-translated into English, although that is possible. The SRT file may either be based on a transcript prepared by a real-time captioner on-site or it could be based on a verbatim transcript prepared offline.

We emphasize that the event types outlined above refer to communicative *events* and not entire videos. Some videos can include instances of each event type, such as a video of an interpreted panel discussion with both deaf and hearing members. In this case, the source language alternates between ASL and English.

4.2.2 Classification of the Corpus

A full discussion on the classifications of parallel corpora is beyond the scope of this paper. There are several. However, we will throw a proverbial wrench into the mix. In broad strokes, according to Fantinuoli and Zanetti (2014), as cited in Wehrmeyer (2019), a corpus is classified with regard to the number of languages it represents (1 = monolingual, 2 = bilingual, 3 or more = multilingual), architecture (comparable or parallel), purpose (general or specialized), modality, and directionality.

In terms of modality, we must consider the visual nature of signed languages and the need for conventions to annotate them so they are machine-readable. We are no strangers to multimodal corpora. In our case, echoing the advice of Heßmann and Vaupel (2008), at some point we will have to decide what consideration we want to give the spoken English lurking within our data. Is a rough transcript sufficient? Is a phonetic transcription merited? What degree of "verbatim" is needed? Moreover, multimodality should also take into account co-speech gestures produced by hearing people who are visible in the videos.

However, the issue of modality is not the wrench. The wrench was alluded to earlier in our discussion in section 4.1, on the ability to access the ASL only by searching through English (at least until we are able to annotate the ASL). This is not to be confused with the directionality of a corpus (unidirectional or bidirectional). In a unidirectional corpus, the translations occur from a source language to a target language. In a bidirectional corpus, the translations occur in both directions. Searches, however, can still be conducted in either language. What do we call a bilingual, multimodal corpus that is only machine-readable in one of its languages? Undoubtedly, in this case, it is an asymmetrical one favoring the majority language with a written system.

4.2.3 Classification of Metadata

Because the GUDA project utilizes Gallaudet's diverse video collection that was originally recorded for numerous reasons, measures are being taken to gather metadata that is useful to corpus research, such as the IMDI initiative and elaborative considerations presented during the ECHO workshop (e.g., Crasborn & Hanke, 2003). Additionally, because the videos were not originally collected for the purpose of being included in a corpus, GUDA researchers are also engaging in re-consent measures, as considered by others (e.g., Chen Pichler, Hochgesang, Simons, & Lillo-Martin, 2016).

The parallel corpus has additional considerations as well. In traditional corpora, a "free translation" is created by an annotator who is not considered a primary participant. However, the interpretations and translations in the parallel corpus were not created by researchers but by participants in the original communicative event. For this reason, participants may or may not be visible in the recordings. That is, in the case of a presentation, the presenter as well as the interpreter, and, if present, the real-time captioner, are all participants and are all providing analyzable utterances in various modalities.

As discussed, annotators are typically not considered participants. However, the work of an annotator has potential as another data point, in which case the "free translation" tier provided by an annotator would have non-traditional benefit within a parallel corpus. Basic identifying information is collected on annotators, but a parallel corpus seaking to analyze the resulting translations may need to consider metadata on par with that collected for any other participant.

4.3 Data Quality of SRT Files

A closer look at the data showed some SRT files were either scarce, partially incomplete, or significantly misaligned. To resolve SRT files with scarce, unuseable information, we recommend including a step to eliminate files below a certain size from the process. Partially incomplete SRT files may be difficult to detect. At this time, we do not have a recommendation for how to eliminate or correct them, other than manual inspection.

Some subtitles in SRT files are significantly misaligned, usually due to errors in timestamping. One possible solution is to add code to check whether the beginning and ending timestamps of an SRT file fall within the duration of its respective video.

4.4 Alignment

Alignment is a significant issue in creating parallel corpora, and it is especially difficult with sign language corpora. Both Meurant, Cleve, and Crasborn (2016) and Wehrmeyer (2019) report on the difficulties in aligning translations with their signed segments.

In Event Type A (ASL) events, SRT files typically segment translations into what can fit and be comfortably read on a screen at the time of its corresponding ASL utterance. While segmentation in traditional sign language corpora is concerned with linguistic boundaries (e.g., Ormel and Crasborn, 2011), segmentation in SRT files is bound more by technological and pragmatic constraints.

In Event Type B (ASL>English) and C (English>ASL) the issue of segmentation is compounded by latency effects. In Event Type B events, the onset of the English translation will typically occur some brief amount of time after the ASL utterance; similarly, in Event Type C events, the onset of the ASL translation will occur some brief amount of time after the English utterance. If real-time transcription services are provided, another level of latency may be introduced between the source and its representation in written English.

Depending on the purpose of the corpus, the latency may be informative in and of itself. For example, in section 2.4 we reported on the study by Cokely (1986), who identified a negative correlation between the duration of the latency and the number of errors committed by the interpreter.

For our initial application of the corpus, which was to find ASL matches used in context for English terms, matches

do not have to be perfectly aligned. Users will likely scroll some time before and after the match to understand the context. Nonetheless, matches must appear within a reasonable window of the segment returned by ELAN.

Other purposes may require a more exact alignment than that provided by SRT files. For instance, data used to train a machine-learning algorithm would have to be well aligned so as not to train the algorithm on the wrong data.

4.5 User Engagement

Many uses of parallel sign language corpora beyond research have been proposed (e.g., Meurant, Cleve, & Crasborn, 2016; Roush, 2016; Wehrmeyer, 2019), yet we do not fully understand the average user's needs, wants, or expectations. Indeed, there is arguably no average user at this point.

We offer the idea that if parallel sign language corpora are to extend beyond the researcher's laboratory, we must investigate users' engagement with the corpora. Therefore, our next steps include gathering reactions from students and professionals to our corpus. Questions to investigate include, for example, how useful is a parallel corpus that can only be searched in one of its languages (e.g., based on English searches)? What features would users want added? What groups of users (translators, interpreters, language learners, educators, etc.) find the corpus most beneficial? Can users think of other implementations of the corpus the researchers have yet to identify?

4.6 Open Access and Sustainability

Currently, the corpus is stored on a hard drive and on three computers in a lab housed by the Department of Interpretation and Translation at Gallaudet University. The next step would be to make this corpus available online so that it is citable and others can benefit from it (e.g., Berez et al, 2018). But this is an issue that other interpreting corpora have. Once the GUDA project identifies a suitable digital location, the data (including the parallel corpus described here) will be made available. We are exploring the possibility of maintaining a dynamic site which we would continuously update with our ongoing work. In addition, we would periodically archive our data (most likely, on a triennial basis) with reputable language archives such as The Language Archive. Stable archived data will ensure access to the corpus beyond the researchers' time at Gallaudet.

4.7 Ethics of Mining Existing Videos

Along with the issue of citability and open access of our parallel corpus is the issue of ethics. By that, we are referring to the need to respect the privacy of the people in the data (their images, voices, and any other identifiable information). Our data comes from existing online videos. Currently there is little consensus regarding treatment of such data—usually because most of those have been based on wholly written texts. The situation is different when it comes to signed languages because we cannot avoid picturing people when we represent their language use. Because our data contains English data, we are including voices of people as well. Management of this kind of potentially identifying data needs to be considered. We will attempt to re-consent all videos by contacting people who are included in the videos (much like outlined in Chen Pichler et al., 2016). Given the immense logistics of such an endeavour, however, one other ongoing solution we are test-driving is an "opt out" mechanism to be provided with each video. We will defer to the preferences of those who appear in the data.

4.8 Growth and Refinement

As stated earlier, the quality of the SRT files was sometimes less than desirable, and it skewed our corpus data. Scripts were written to handle the data at hand and were therefore useful to construct only this corpus. Moreover, since the construction of this corpus in November 2019, we have identified other video sources and collaborators.

Therefore, a future iteration of this parallel corpus would include writing more general scripts designed to work with any incoming data and that can be shared with the wider research community. In addition, procedures would be put in place to handle and flag the issues identified in this paper (e.g., removing or repairing corrupt SRT files). With new video sources and a more streamlined process for categorizing the videos, we could also investigate the feasibility of creating specialized corpora.

5. Conclusion

This is a first attempt at creating a sizable parallel corpus, albeit only searchable through one of its languages, by leveraging SRT files. Lessons learned and considerations outlined in this paper may serve as a blueprint for future endeavors and other scholars.

6. Bibliographical References

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and technology: In honor of John Sinclair* (pp. 233–252). John Benjamins.

Baker, P., Hardie, A., & McEnery, T. (2006). *A glossary of corpus linguistics.* Edinburgh University Press.

Berez-Kroeker, A. L., Gawne, L., Kung, S. S., Kelly, B. F., Heston, T., Holton, G., Pulsifer, P., Beaver, D. I., Chelliah, S., Dubinsky, S., Meier, R. P., Thieberger, N., Rice, K., & Woodbury, A. C. (2018). Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics*, 56(1), 1–18.

Chen Pichler, D., Hochgesang, J., Simons, D., & Lillo-Martin, D. (2016). Community Input on Re-consenting for Data Sharing. In E. Efthimiou, S.-E. Fotinea, T. Hanke, J. A. Hochgesang, J. Kristoffersen, & J. Mesch (Eds.), Workshop Proceedings: 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining / Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016) (29-34). Paris: European Language Resources Association (ELRA).

Cokely, D. (1986). The effects of lag time on interpreter errors. *Sign Language Studies, 53*, 341–375. https://doi.org/10.1353/sls.1986.0025

Crasborn, O., & T. Hanke. (2003, May 8–9). (version Jan 2010). Metadata for sign language corpora. Background document for an ECHO workshop, Radboud University Nijmegen.

Fantinuoli, C. & Zanettin, F. (2014). Creating and using multilingual corpora in translation studies. In C. Fantinuoli & F. Zanettin (Eds.), *New directions in corpus-based translation studies* (pp. 1–10). Language Science Press.

Frishberg, N. (2010, December 3). *Repurposing corpus materials for interpreter education* [Workshop presentation]. Sign Linguistics Corpora Network, Berlin, Germany. https://www.ru.nl/publish/pages/607111/slcn4_frishberg.pdf

Gallaudet University. (2007, November). *Mission and Goals. https://www.gallaudet.edu/about/planning-for-the-future/mission-and-goals*

Hellwig, B., Van Uytvanck, D., … & Geerts, J. (2019). *ELAN - Linguistic annotator* (Ver. 5.8). The Language Archive. https://www.mpi.nl/corpus/manuals/manual-elan.pdf

Heßmann, J., & Vaupel, M. (2008). Building up digital video resources for sign language interpreter training. In O. Crasborn et al. (Eds.), *3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora* (pp. 74–77). European Language Resources Association.

Hochgesang, J. A., Crasborn, O., & Lillo-Martin, D. (2018). Building the ASL Signbank: Lemmatization principles for ASL. In M. Bono et al. (Eds.), *8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community* (pp. 69–74). European Language Resources Association.

Hochgesang, J. A., Willow, J., Treviño, R., & Shaw, E. (2019, September 26–28). *Gallaudet University Documentation of ASL (GUDA) - Whither a corpus for ASL?* [Poster presentation]. 13th Conference of Theoretical Issues in Sign Language Research (TISLR13), Hamburg, Germany. https://doi.org/10.6084/m9.figshare.9842696.v1

McEnery, T., & Hardie, A. (2012). *Corpus linguistics.* Cambridge University Press.

Meurant, L., Cleve, A., & Crasborn, O. (2016). Using sign language corpora as bilingual corpora for data mining: Contrastive linguistics and computer-assisted annotation. In E. Efthimiou et al. (Eds.), *7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining* (pp. 159–166). European Language Resources Association.

Ormel, E., & Crasborn, O. (2011). Prosodic Correlates of Sentences in Signed Languages. A Literature Review and Suggestions for New Types of Studies. *Sign Language Studies*, 12(2), 279–315.

Shlesinger, M. (1998). Corpus-based interpreting studies as an offshoot of corpus-based translation studies. *Meta, XLIII* (4).

Roush, D. (2016, October 26–29). *Learning benefits of a translation corpus for novice ASL-English interpreters* [Poster presentation]. 2016 Biennial Conference of the Conference of Interpreter Trainers, Lexington, KY.

Wehrmeyer, E. (2019). A corpus for signed language interpreting research. *International Journal of Research and Practice in Interpreting, 21*(1) 62–90. https://doi.org/10.1075/intp.00020.weh

7. Language Resource References

Crasborn, O., Zwitserlood, I., & Ros, J. (2008). The Corpus NGT. An open access digital corpus of movies with annotations of Sign Language of the Netherlands. Centre for Language Studies, Radboud University Nijmegen, ISLRN 175-346-174-413-3.

Hochgesang, J.A., Crasborn, O., & Lillo-Martin, D. (2020) ASL Signbank. New Haven, CT: Haskins Lab, Yale University. https://aslsignbank.haskins.yale.edu/

Meurant, L. (2015). Corpus LSFB. First digital open access corpus of movies and annotations of French Belgian Sign Language (LSFB). Laboratoire de langue des signes de Belgique francophone (LSFB-Lab). FRS-F.N.R.S. et Université de Namur.

Author Index