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

# Workshop on Legal and Ethical Issues in Human Language Technologies (LEGAL2020)

# PROCEEDINGS

Khalid Choukri (ELRA/ELDA, France), Kirster Linden (University of Helsinki, Finland), Mickaël Rigault (ELRA/ELDA, France), Ingo Siegert (Otto von Guericke University, Magdeburg, Germany) (eds.)

# Proceedings of the LREC 2020 Workshop on Legal and Ethical Issues in Human Language Technologies (LEGAL2020)

Edited by: Khalid Choukri, Kirster Linden, Mickaël Rigault, Ingo Siegert

**ISBN: 979-10-95546-37-5**
**EAN: 9791095546375**

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

# Introduction

Welcome to the LREC2020 Workshop on "Legal and Ethical Issues in Human Language Technologies".

First of all we would like to thank all the persons who contributed to the achievement of this Workshop in spite of the special circumstances that compel us to hold this event in a different way than we usually do. We would also like to send a thought to all essential personnel that helped all of us continue our tasks during the lockdown period, and to all the victims of this pandemic. In line with previous workshops, this years' papers are focused on the application of privacy regulations both in Europe and abroad. There is still great development, not so much with regards to the texts themselves, but with regards to their implementation. The practices of research teams and the further definitions of the legal framework by judges and lawmakers have helped to further understand the principles of privacy and the definitions of best practices for ethical research in the field of language technology. These papers reflect on the evolution of the regulatory framework and its impact on the research activities in the Human Language Technology community.

The Organising Committee

**Organizers:**

**Khalid Choukri (ELDA), France**
**Kirster Linden (University of Helsinki), Finland**
**Mickaël Rigault (ELDA), France**
**Ingo Siegert (Otto von Guericke University, Magdeburg), Germany**

# Table of Contents

*Proceedings of the workshop on Legal and Ethical Issues (Legal2020)*, pages 1–3
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020


# GDPR – A Game Changer for Acoustic Interaction analyzes

**Ingo Siegert**[1], **Vered Silber Varod**[2], **Pawel Kamocki**[3]
[1]Mobile Dialog Systems, Otto von Guericke University Magdeburg Germany
[2]Open Media and Information Lab, The Open University Israel, Israel
[3]Leibniz Institut für Deutsche Sprache, Mannheim, Germany
ingo.siegert@ovgu.de, vereds@openu.ac.il, kamocki@ids-mannheim.de

**Abstract**

Human interaction analyzes are essential to study social interaction, conversational rules, and affective signals. These analyzes are also used to improve models for human-machine interaction. Besides the pure acoustic signal and its transcripts, the use of contextual information is essential. Since the enforcement of the GPDR for the EU in 2018, there has been an increased uncertainty among scientists and participants. The discussion about the EU GDPR raised the awareness of personal rights and personal data recordings. This contribution aims to discuss issues of collecting personal and contextual data during acoustic interaction in terms of scientists' needs and GDPR demands.

**Keywords:** acoustic interaction, contextual data, GPDR, personal data

## 1. Introduction

The General Data Protection Regulation (EU Regulation 2016/679, hereinafter: the GDPR) entered into application on May 25, 2018. The aim of the GDPR is on one hand to unify all data protection laws across the European Union and on the other hand to protect the information about all EU residents against unlawful processing and privacy breaches. The GDPR has raised awareness about privacy-related rights throughout the EU. Data misuse has already led to noteworthy fines, including a 20k EUR fine against a French translation company Uniontrad for videotaping its employees, or against an Italian hospital (Azienda Ospedaliero Universitaria Integrata di Verona) of 30k EUR for not adequately protecting patient personal health records from unauthorized treatment. A danish taxi company stored data from eight million trips and thus violated the minimization principle of the GPDR, resulting in a fine of 161k EUR. The so-far highest fine (204M EUR) was condemned against British Airways due to a cyber incident, where 500,000 customers' personal data were compromised.

These examples of data misuse and the raising awareness have a direct impact on research activities, not only among scientists but also among experimental subjects.

The GDPR regulates the way data can be collected, stored, and processed (analyzed, exchanged, etc. ((Sveningsson Elm, 2009)). This entails the constitution of "personal data" and its efficient anonymization. According to Art. 4, 1. of the GDPR "personal data" is defined as "any information relating to an identified or identifiable natural person" ("data subject"). This concept of 'personal data' is significantly broader than the concept of 'personally identifiable information' (PII) used e.g. in the US. Personal data definition.

In contrast, interaction analyzes require huge data collections together with contextual information of the participants, in order to understand the interaction process, the individual behavior and develop proper models (Dudzik et al., 2019). These needs are challenging regarding the GDPR, leading to a huge uncertainty for data collection and data sharing activities.

Up until now, scientists tried to deal with it on their own and to help their peers by publishing documents and papers on ethical issues. Batliner and Schuller (2014), for example, list crucial ethical issues, including the challenge to guarantee the consent and the privacy of the subjects and the need to encode the data to guarantee this privacy (Batliner and Schuller, 2014).

This contribution aims to discuss the above issues by highlighting needs scientists have for analyzing interactions by giving examples in which additional "personal data" are needed but their storage and exchanging is crucial according to the GDPR.

## 2. Examples of the need for "personal data" in interaction analyzes

An important aspect of human perception is the processing of additional contextual information (Dudzik et al., 2019; Truong et al., 2007). The same holds true for technical systems. They must implement these human abilities and analyze human interaction signals together with additional contextual information. Therefore developers need databases capturing the context of interactions as well as the behavior expressed in them, which is also denoted as enriched data (Böck et al., 2019). Recent literature already surveys empirical research on how the decoding of behavioral signals in emotion perception benefits from contextual information (Wieser and Brosch, 2012) and how perceivers make use of contextual knowledge in interpreting affective behavioral signals (Aviezer et al., 2017). Important contextual categories are developed in (Dudzik et al., 2019), comprising age, gender, cultural embedding – nationality and ethnic background, language, and occupation. Furthermore, also personality traits, as NEO-FFI or SVF, and additional measurable signals are helpful, as it will be shown in the following.

Some examples where contextual data are needed to improve the interaction analyzes and modeling will be shortly

discussed in the following. Age and gender information of a user is particularly required to improve automatic emotion recognition due to speaker-group dependent models (Siegert et al., 2014c) and could even improve multimodal recognition for fragmentary data (Siegert et al., 2013). For example, in order to improve the emotion recognition significantly, recognition models make use of factors that affect the vocal tract, such as aging-effects, to improve their acoustic models by personalizing it to a specific age group. Also for human annotation of conversations not only the speech is important but also facial information (Siegert et al., 2014a), which allows an improvement of the identification of the participant.

Only a few analyzes so far deal with personality traits as additional contextual information. Although it is known that certain personality traits play an important role in communication (Cuperman and Ickes, 2009; Funder and Sneed, 1993; Weinberg, 1971). In (Gossen et al., 2017) the authors showed that the Incorporation of the contextual information on the personality trait "extraversion" improves the long term modeling of interactions. Furthermore, it is shown that information about the stress-coping ability of participants is useful to link exhaustive filled pause usage for the detection of challenging tasks (Siegert et al., 2014b). These examples underline the importance of both the acoustic signal and the contextual information (metadata) of the subjects to acoustic interaction research. Especially for automatic affect recognition systems the incorporation of metadata is beneficial.

## 3. GDPR-issues of recording contextual data

Recording contextual data of a participant can, even if all direct identifiers (name, birth, residence) are deleted, be used to identify a specific participant. A participant is identified when it's singled out from a group, typically by a sufficiently unique name-surname combination, but other identifiers (e.g., username or ID number, or in a certain context – a photograph) can also be taken into account. Moreover, a person is 'identifiable' if it can be singled out from a group by any means reasonably likely to be used (such as cross-referencing with data from social networks). Many examples are known, that not much data is generally needed to identify a person, even if the records are anonymized; e.g. the combination of zip codes, birth date and sex from anonymized data together with voter databases is enough to identify individuals (Ohm, 2010) Or that for identifying users of a famous video-streaming platform using knowledge about some movie ratings (Narayanan and Shmatikov, 2008).

What does that mean for the recording of contextual data? According to art. 5.1, c) of the GDPR, personal data should be 'adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed'. The development of an experimental design and the conduction of the experiment can be very elaborate. Especially for interaction analyzes, as most of them are fundamental research, at the beginning of the experiment it is not clear which contextual factors are relevant for the task. Sometimes, additional questions arise during the evaluation of the data or new collaborative ideas are showing up during the presentation of the analyzes results. Thus, it is not always known nor desirable to limit the recording of personal data for research analyzes. This principle (which existed also under the 1995 Data Protection Directive), referred to as 'data minimization', is arguably the biggest hurdle for data-intensive research and technology especially as the GDPR does not allow any derogations from the principle for research purposes.

Furthermore, as the examples in the introduction showed, the protection of personal data against cyber incidents is crucial. Researchers mostly do not have the capacity nor the knowledge how to properly secure the data against misuse and on the same time still allow access to the data for authorized persons.

In this context, the German Research Foundation (DFG) for example explicitly encourages applicants to request funding for the preparation of research data for subsequent reuse or transfer. But this mostly covers the preparation, long-term archiving, and accessibility of data. Aspects relating to the compliance with the GDPR (access control, anonymization techniques, selective data access) are not in the focus so far.

## 4. Conclusion

The new regulations pose a new situation where despite not sharing personal data, or even not collecting at all, there is no proposed solution at the moment, at least for acoustic interaction research. Researchers might need a combined policy of legal and academic authorities.

One possibility is that the research community be more thorough, by disconnecting the assignment of context data to certain persons. This can be done by using ranges or broader classes for contextual data. For example, in (Silber-Varod et al., 2019) they used solely the acoustic signal and speaker-sex attribute, as the data was proprietary by an industrial company.

Another possibility is that recorded data is anonymized – hence the importance of anonymization or data omission for research activities gets important – as well as a proper access control infrastructure still allowing the share research data has to be developed. Hereby, it has to be noted that already the voice recordings itself reveal the speakers' identity. This constitutes a bigger challenge to cope with, especially in terms of anonymization.


Aviezer, H., Ensenberg, N., and Hassin, R. R. (2017). The inherently contextualized nature of facial emotion perception. *Current Opinion in Psychology*, 17:47 – 54. Emotion.

Batliner, A. and Schuller, B. (2014). More than fifty years of speech and language processing - the rise of computational paralinguistics and ethical demands. In *ETHICOMP 2014, 25-27 June 2014, Paris, France*.

Böck, R., Egorow, O., Höbel-Müller, J., Requardt, A. F., Siegert, I., and Wendemuth, A., (2019). *Anticipating the User: Acoustic Disposition Recognition in Intelligent Interactions*, pages 203–233. Springer International Publishing, Cham.

Cuperman, R. and Ickes, W. (2009). Big five predictors of behavior and perceptions in initial dyadic interactions:

Personality similarity helps extraverts and introverts, but hurts 'disagreeables'. *J Pers Soc Psychol*, 97:667–684.

Dudzik, B., Jansen, M., Burger, F., Kaptein, F., Broekens, J., Heylen, D. K. J., Hung, H., Neerincx, M. A., and Truong, K. P. (2019). Context in human emotion perception for automatic affect detection: A survey of audiovisual databases. In *Proc. of 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 206–212, Sep.

Funder, D. C. and Sneed, C. D. (1993). Behavioral manifestations of personality: An ecological approach to judgmental accuracy. *J Pers Soc Psychol*, 64:479–490.

Gossen, T., Siegert, I., Nürnberger, A., Hartmann, K., Kotzyba, M., and Wendemuth, A., (2017). *Modeling aspects in human-computer interaction - adaptivity, user characteristics and evaluation*, pages 57–78. Springer International Publishing, Cham.

Narayanan, A. and Shmatikov, V. (2008). Robust deanonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125, May.

Ohm, P. (2010). Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review*, 57:1701–1777.

Siegert, I., Glodek, M., Panning, A., Krell, G., Schwenker, F., Al-Hamadi, A., and Wendemuth, A. (2013). Using speaker group dependent modelling to improve fusion of fragmentary classifier decisions. In *Proc. of 2013 IEEE CYBCONF*, pages 132–137, Lausanne, Switzerland.

Siegert, I., Böck, R., and Wendemuth, A. (2014a). Inter-Rater Reliability for Emotion Annotation in Human-Computer Interaction – Comparison and Methodological Improvements. *Journal of Multimodal User Interfaces*, 8:17–28.

Siegert, I., Haase, M., Prylipko, D., and Wendemuth, A. (2014b). Discourse particles and user characteristics in naturalistic human-computer interaction. In Masaaki Kurosu, editor, *Human-Computer Interaction. Advanced Interaction Modalities and Techniques*, volume 8511 of *LNCS*, pages 492–501. Springer, Berlin, Heidelberg, Germany.

Siegert, I., Philippou-Hübner, D., Hartmann, K., Böck, R., and Wendemuth, A. (2014c). Investigation of speaker group-dependent modelling for recognition of affective states from speech. *Cognitive Computation*, 6(4):892–913.

Silber-Varod, V., Lerner, A., Carmi, N., Amit, D., Guttel, Y., Orlob, C., and Allouche, O. (2019). Computational modelling of speech data integration to assess interactions in b2b sales calls. In *IEEE 5th International Conference on Big Data Intelligence and Computing (IEEE DataCom 2019)*, pages 125–127.

Sveningsson Elm, M., (2009). *How do various notions of privacy influence decisions in qualitative internet research?*, pages 69–87. SAGE Publications, Thousand Oaks.

Truong, K. P., van Leeuwen, D. A., and Neerincx, M. A. (2007). Unobtrusive multimodal emotion detection in adaptive interfaces: Speech and facial expressions. In Dylan D. Schmorrow et al., editors, *Foundations of Augmented Cognition*, pages 354–363, Berlin, Heidelberg. Springer Berlin Heidelberg.

Weinberg, G. M. (1971). *The psychology of computer programming*. Van Nostrand Reinhold, New York, USA.

Wieser, M. and Brosch, T. (2012). Faces in context: A review and systematization of contextual influences on affective face processing. *Frontiers in Psychology*, 3:471.

*Proceedings of the workshop on Legal and Ethical Issues (Legal2020)*, pages 4–8
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020


# GDPR Compliance for task-oriented dialog systems conception

**Leon-Paul Schaub**[1],**Christine Bruzaud**[2], **Patrick Paroubek**[3],
LIMSI-CNRS[1 3], Akio[1 2], Universite Paris-Saclay[1 3]
Campus Universitaire batiment 507, Rue John Von Neumann 91400 Orsay [1 3]
43 Rue de Dunkerque, 75010 Paris[1 2]
Espace Technologique Bat. Discovery - RD 128 - 2e et, 91190 Saint-Aubin [1 3]
schaub@limsi.fr[1], cbruzaud@akio.com[2], pap@limsi.fr [3]

**Abstract**

We address the issue of applying the recent General Data Protection Regulation when designing or deploying goal-oriented dialog systems (task-oriented dialog systems). This implies answering questions like who among the actors involved is responsible of the data control during the interactions between a bot and a user, who shall manage the data transfer, storage and future access/modification requests. To answer all these questions, we propose a protocol for the GDPR-compliant task-oriented dialog system conception checking called GCCP to provide guidelines for both scientific research and industrial deployment.

**Keywords:** Goal-oriented dialog system, GDPR, task-oriented dialog system's conception, Data Management Plan

## 1. Introduction

In France, personal data protection law is not a new idea (Piolle and Demazeau, 2008). In 1978, the Informatics and Liberty law (LIL) was voted. At the same time the law enforcement Informatics and Liberties National Committee (CNIL) was created. However, in the last fifteen years and the rise of social networks, numerical technology revolutionized both people's everyday life[1] and companies' business practices (Bonchi et al., 2011). This is why in 2016 the EU Parliament voted the GDPR to protect individual privacy and prevent misuse of personal information. Here are the main evolution brought by this new law:

- Transparency becomes an obligation (Goddard, 2017)
- Responsibilities are re-balanced (Lindqvist, 2017)
- New concepts are created or instantiated: profiling, right to be forgotten, privacy by design. (Spiekermann, 2012)

The artificial intelligence behind text mining techniques is analytical : it takes data as input and according to all the texts the AI has seen before, it applies an algorithm (classification, translation, parsing...) depending on the task(s) it has been created for. However our studies focus on a technology that uses not only analysis, but also human-machine interaction (HMI) : dialog systems and more precisely task-oriented dialog systems (tods). It is a computer program built to interact with a human in order to complete a specific task, like booking a hotel, buying clothes online or answering questions about a particular device, system or service. A survey on this topic was written by (Schaub and Vaudapiviz, 2019). The problem of the task-oriented dialog systems with GDPR and data management is the real-time interaction. Indeed, whether the text mining task is sentiment analysis, dependency parsing or question-answering, the personal data anonymization is not the same issue to achieve good performance, because the AI does not need to have any kind of interaction with the user. The main difference with the dialog task is the need for the task-oriented dialog system to be empathic to improve human acceptance. (Tahara et al., 2019) improve user satisfaction by learning emotion embeddings to have a better human understanding. In the next sections we will provide some detailed elements of data management (Kamocki et al., 2018) in order to create a protocol to check GDPR compliance during task-oriented dialog systems construction. We will also explore related works on GDPR compliance for HMI and finally suggest future experiments to evaluate the robustness of the proposed protocol.

[1]http://www.comonsense.fr/influence-medias-sociaux-vie-quotidienne/

## 2. Data in dialog systems

In this section, we will define the technical issues of a GDPR-compliant's task-oriented dialog system and address the problem of dialogue data management. Finally, we will discuss the problem of anonymization with real-life cases.

### 2.1. task-oriented dialog system architecture

In this paper, we consider a task-oriented dialog system as a text-driven G-O dialog system. A task-oriented dialog system's purpose is to understand the users intention, optimize its internal representation of the user's goals and its own desire during conversation (subgoals). Althought there exists many possible architecture for dialog systems, as described in (Schaub and Vaudapiviz, 2019), a common architecture (Young et al., 2012) of a task-oriented dialog systems has three main components (Figure 1) :

**a. Natural Language Understanding NLU** parses user new input and encodes it in its internal memory under the form of slots or frames like a dictionary that is updated af-

ter each speaking turn (El Asri et al., 2017). In this component, as the input comes directly from the user, there might be personal data.
**b. Dialogue Manager DM** explores the updated dictionary and according to its long term memory, under the form of a model of language learnt from all the past conversations, and an external knowledge base, it tracks the dialogue state to decide what answer needs to be outputted (Madotto et al., 2018). During the transformation step, the personal data is part of the internal representation and thus as we explain in 2.2, can be used to retrieve the information from the long-term memory in order to output the right answer.
**c. Natural Language Generation NLG** transforms (decodes) the answer decision from the DM into natural language output under the form of templates in a retrieval-based generator (Wu et al., 2019) or with generative-based generator (Serban et al., 2017; Li et al., 2017). Depending on what has been learnt previously, there might be personal data output as well.

## 2.2. The problem of anonymization

As the GDPR started to be applied last year, many companies and even research laboratories working on text data focused their work on finding the best way to anonymize documents. (Di Cerbo and Trabelsi, 2018) propose an overview of classic techniques of text anonymization and a novel approach based on state-of-the-art machine learning algorithms. (Kleinberg et al., 2018) developed an open-source named-entities anonymizer software called NETANOS. More recently (Kim et al., 2019) introduce a protocol to properly anonymize the data, to be totally GDPR-compliant showing improvements of the anonymization techniques. However, as explained by (Bottis and Bouchagiar, 2018) it is very hard, probably impossible to perfectly anonymize all personal due to constant improvements of re-identification techniques and thus the need of periodically make evolve the anonymizer (Hayes et al., 2017).
Once again let us assume that there exists a perfect anonymizer. Indeed, a task-oriented dialog system fed with anonymized input is GDPR-compliant, but then it loses the capacity of remembering crucial information during a goal-oriented conversation such as who it is talking to.
For the learning phase, where the AI behind the task-oriented dialog system learns from past conversation, anonymization is not an issue, as long as the original conversation structure is kept, in order to be similar to the real conversation the task-oriented dialog system will have to face during the deployment/evaluation phase. The anonymization could be problematic though for new conversations as we know that one of the main condition for a machine to be human-friendly is to be human-like (Ouali et al., 2019), and we doubt that having an amnesic task-oriented dialog system is a way to achieve human-likelihood and empathy simulation. As we explained earlier, a good employee needs to show empathy during the dialogue so the customer satisfaction probability is increased.
Let us assume now that the costumer does not care during a conversation whether the task-oriented dialog system shows empathy or not. There are at least two scenarios where a complete data anonymization remains a problem.

### 2.2.1. Customer recall

Imagine the situation when a customer C after ending a conversation with an agent A, calls some time later, for any good reason, the same service and it is the same agent who answers the call. In a normal situation, if the two calls are made within the same hour, C expects A to remember the call or at least some piece of information related to it such as : the reason of the first call, the name of C, and eventually the most salient problems faced. In most cases, C's satisfaction will be correlated with A's recall's capacities. Even if a task-oriented dialog system B is well trained on what we named the first call, it might face difficulties to satisfy the customer on second call conversation. There is an imbalance between C's expectations and B's capacities. Even though on the first call, C did not need any empathy signs from B, on second call it will be different because a bond already exists between C and B from C's point of view. But because of the anonymization, even if B understands that it is a second call situation, it will never be able to recognize C as the author of a previous call.

### 2.2.2. Personal data recurrence

This second situation is not a definitive handicap as the previous one but the task-oriented dialog system technology would make an improvement if it had a solution to the situation.

Imagine the situation when an Internet service provider receives thousands of calls on the same day because there is a huge breakdown in a specific area. After several calls from frustrated customers, when a new customer C calls with the same tone or writes an email with similar semantics that previous ones, an agent A knows without even asking what is all about : the breakdown, the place where it happened, and even C's complains. This inference capacity helps A to be more efficient during the new conversation and provide to C all the needed information. Moreover, A knows how to calm down C after experimenting techniques with previous unhappy customers all day long. Now, if the agent is instead our task-oriented dialog system B, this one-day-only improvement is impossible due to anonymization : in the GDPR it is explicitly said that any information that can identify a person shall be transformed. This included customer's location and emotional state. Therefore there is no way that B, even if it had a one-day memory, could connect previous complains with C's. In B's memory, it will be an astonishing coincidence that the same breakdown occurs so many times this day.

This is why in section 3 we introduce a protocol that could help improving task-oriented dialog systems capacities while remaining GDPR-compliant.

## 3. The GCCP : GDPR Compliance task-oriented dialog system Protocol

Here we describe the protocol for task-oriented dialog system conception through the pipeline illustrated in the Figure 2.

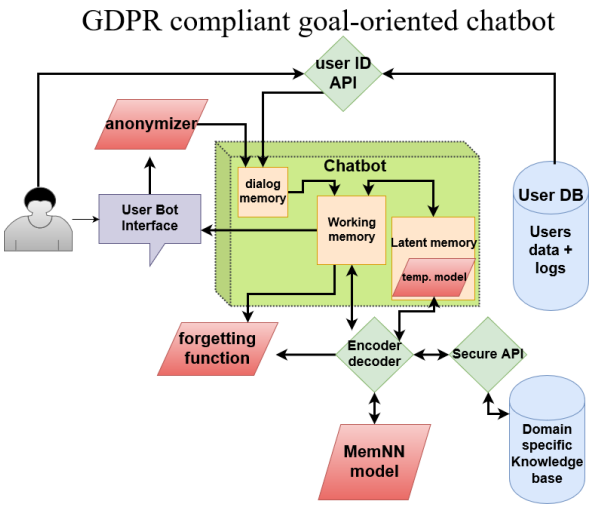


Figure 1: GDPR compliant task-oriented dialog system

### 3.1. The task-oriented dialog system's conception

Here are the main components of the task-oriented dialog system architecture :

**a. User-bot interface UBI** : it can be a chat box like messenger or a tool integrated into the CRM platform

**b. Anonymizer** : an external module used to anonymize a user's input at each speaking turn. It can be trained or symbolic

**c. User database and API** : it is the CRM database used to retrieve the user's profile data if the bot needs it.

**d. Dialog memory** : it is the first and smallest of the three internal modules of the task-oriented dialog system, it's purpose is to preprocess anonymized input into a dictionary form. It is wiped after each speaking turn.

**d. Working memory WM** : it is the second module of the task-oriented dialog system and the core of the process because this is were the memories from the conversation are enhanced by the Memory network dialogue model (h.) and the external database (i.) but also by the latent memory (e.). Once the conversation is over, the WM forgets (g.) the personal data retrieved at the beginning of the conversation and encodes the conversation representation plus the dialog result to the latent memory. It is also in this module that the output is generated and sent to the UBI at each speaking turn. The WM is limited and the older conversations within it are wiped after some time to stock the new ones.

**e. Latent memory** : it is the third module of the task-oriented dialog system and it contains all the conversations the WM wiped but without the personal data. Its capacity is also limited but much bigger that working memory. However it has not an active processing function. When its size reaches some milestone, it learns a (neural) model of the conversations of the day plus the dialog result. During a conversation, it is used by the WM as the competing information retrieval source of the MNDM. It is wiped at the end of the day.

**f. Temporary model** It is the model learnt by the Latent memory after it has enough conversations to do so. At the end of the day, the model is encoded into the MNDM to improve it, and then wiped.

**g. Forgetting** : it is a learnt function used twice during the process : first at the end of a conversation to purge any personal data left in the WM before it connects to the latent memory, and second at the end of the day to remove any irrelevant information or to check if no personal data is left in the temporary model data.

**h. Memory network dialogue model MNDM** : it is the model representing all the past conversations (dialogue corpus) learnt. It is the task-oriented dialog system's long-term memory. In the architecture it is an external model in case that for any reason the task-oriented dialog system needs to switch to a different behaviour than the one learnt by the model. It is inspired from (Zhang et al., 2019)

**i. External knowledge base EKB** : it is the information system provided by a domain client such as product list or official documentation.It represents the semantic memory of the task-oriented dialog system and shall be disconnected from the MNDM because the same MNDM can be used for different EKB and to avoid that the task-oriented dialog system becomes too domain specific.

As was shown in the section 2, due to the opacity of state-of-the-art models in dialog systems, it might be difficult to build a fully end-to-end architecture, for security reasons despite their advantages such as training speed and model size (Rajendran et al., 2018; Rajendran et al., 2019). However, what we call the task-oriented dialog system's long term memory, which represents the neural model learnt from past conversations can be an end-to-end system (Wu et al., 2018). In our architecture, the task-oriented dialog system itself does not contain the long-term memory, neither the anonymizer tool, nor the external domain specific knowledge base.

### 3.2. Define a compliant GCCP

0. The first step, not the least important is to **ask the users if they accept** that the data during the conversation may be used afterwards to improve the task-oriented dialog system.

1. As we saw in section 2, **the anonymization** is necessary step in the task-oriented dialog system's conception. It is named privacy by design. The anonymizer shall be called for each user's input.

2. However, if the task nature needs some personal data such as an email in order to identify the user's file or account, or boarding pass.. The task-oriented dialog system should be able to **retrieve this information from**

**the user's database**. To do so, the architecture must implement an API with a temporary User ID to provide the task-oriented dialog system all the information it needs to fulfill its purpose.
3. **The UID is stocked** in the task-oriented dialog system's dialogue memory during the first speaking turn and sent to the WM.
4. The anonymized **input is then encoded in the MNDM module** and **the latent memory.** During the decoding, an API call is made to the external knowledge in case of it is necessary for the conversation or if an API call has to be made. The result of the decoding is the ouput of the speaking turn. The **process is repeated during all the conversation.**
5. At the end of the conversation, **the UID is kept in the WM** during a few minutes (up to one hour) in case the same user is engaging a new conversation during this time.
6. After these minutes, **forgetting function is then called** in order to remove from WM all personal data. It is **stocked in the latent memory** representing the daily conversations.
7. When the latent memory starts to get bigger, a **model of the daily conversation** is learnt by the task-oriented dialog system, to know if there are particular trends this day that should be salient for the task-oriented dialog system WM.
8. At the end of the day, the **one-day model is encoded into the long term memory**, and the forgetting function is called, in case that personal data unfortunately remained in the latent memory. **The latent memory is deleted.**
9. Finally, the **MNDM is retrained** with the new conversations of the day.
By following this protocol, the task-oriented dialog system is both GDPR-compliant and efficient for any task.

### 3.3. Limits of the GCCP

Althought the GCCP seems to obey to GDPR rules, they are several limits that must be noticed.

As we said in section 2, there is not a perfect anonymizer, and even if new models become very accurate, there might have some information that avoid the anonymization.
Second, the language model where previous conversations are stocked is also learnt, and therefore may also contain personal data. When in many cases, adding new conversations will improve the model efficiency, it may also increase the danger of personal data being output during the inference.
Finally, as the task-oriented dialog system is available online, there might be a security issue when it makes API calls to the user's database. A study needs to be made in order to verify if the security danger is real or not.

### 3.4. GDPR compliance

According to GDPR official checklist [2] inspired from [3] we attempted to provide the seven requirements in order to be GDPR-compliant.
**1. Optaining consent** : it corresponds to the step 0 of the GCCP.
**2. Timely breach notification** : we have 72 hours to report a data breach. As the personal data is deleted within the hour, the risk of data breach is very limited.
**3. Right to access data** : any customer is allowed to access the data collected about him/her. This is not a problem as an API exists between the customer and the user database, independently of the task-oriented dialog system.
4. **Right to be forgotten** : the customers can request whenever they want that any information concerning then is deleted. The task-oriented dialog system only learns anonymized conversations and the personal data is deleted within the hour (or even before) from the task-oriented dialog system's WM.
**5. Data portability** : users can optain all the data they transmitted to reuse it outside the company. Once again, the task-oriented dialog system does not keep this information, so it is "safe" from this requirement.
**6. Privacy by design** : The system shall be design with proper security protocols. As the task-oriented dialog system "outsources" many of its functions, the risk lowers because when a failure is noted, it is much easier to detect it an shut it if it is outside the task-oriented dialog system in a well identified module.
**7. Potential data protection officers** : this forces a company or an organisation suchas a research lab to appoint a a data protection officer (DPO) to make sure that the previous GDPR requirements are respected. This does not directly depend on the task-oriented dialog system.

## 4. Conclusion

We explained some issues inherent to goal-oriented dialog systems conceptions to be compliant with GDPR. We illustrated with two examples that anonymization can sometimes be a problem to build an efficient task-oriented dialog system but still mandatory to be GDPR compliant. To solve this contradiction, we proposed the GCCP (GDPR compliance task-oriented dialog system protocol) in order to insure a performant task-oriented dialog system by providing the scheme of a fully operational pipeline but still respecting the GDPR requirements. In the future we will test this pipeline with private data but also with public corpora to confirm the robustness of this pipeline inspired by the GCCP.

## 5. Acknowledgements

This work was co-financed by ANRT and Akio under the CIFRE contract 2017/1543

[2] https://gdpr.eu/checklist/

[3] https://www.coredna.com/blogs/general-data-protection-regulation#2

## 6. Bibliographical References


Bonchi, F., Castillo, C., Gionis, A., and Jaimes, A. (2011). Social network analysis and mining for business applications. *ACM Trans. Intell. Syst. Technol.*, 2(3):22:1–22:37, May.

Bottis, M. and Bouchagiar, G. (2018). Personal data v. big data in the eu: Control lost, discrimination found. *Open Journal of Philosophy*, 08:192–205, 01.

Di Cerbo, F. and Trabelsi, S. (2018). Towards personal data identification and anonymization using machine learning techniques. In András Benczúr, et al., editors, *New Trends in Databases and Information Systems*, pages 118–126, Cham. Springer International Publishing.

El Asri, L., Schulz, H., Sharma, S., Zumer, J., Harris, J., Fine, E., Mehrotra, R., and Suleman, K. (2017). Frames: a corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 207–219, Saarbrücken, Germany, August. Association for Computational Linguistics.

Goddard, M. (2017). The eu general data protection regulation (gdpr): European regulation that has a global impact. *International Journal of Market Research*, 59(6):703–705.

Hayes, J., Melis, L., Danezis, G., and Cristofaro, E. D. (2017). LOGAN: evaluating privacy leakage of generative models using generative adversarial networks. *CoRR*, abs/1705.07663.

Kamocki, P., Mapelli, V., and Choukri, K. (2018). Data management plan (DMP) for language data under the new general da-ta protection regulation (GDPR). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Kim, B., Chung, K., Lee, J., Seo, J., and Koo, M.-W. (2019). A bi-lstm memory network for end-to-end goal-oriented dialog learning. *Computer Speech and Language*, 53:217 – 230.

Kleinberg, B., Mozes, M., van der Toolen, Y., and Verschuere, B. (2018). Netanos - named entity-based text anonymization for open science. *OSF*, Jan.

Li, J., Monroe, W., Shi, T., Jean, S., Ritter, A., and Jurafsky, D. (2017). Adversarial learning for neural dialogue generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2157–2169, Copenhagen, Denmark, September. Association for Computational Linguistics.

Lindqvist, J. (2017). New challenges to personal data processing agreements: is the GDPR fit to deal with contract, accountability and liability in a world of the Internet of Things? *International Journal of Law and Information Technology*, 26(1):45–63, 12.

Madotto, A., Wu, C.-S., and Fung, P. (2018). Mem2Seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1468–1478, Melbourne, Australia, July. Association for Computational Linguistics.

Ouali, L. O., Sabouret, N., and Rich, C. (2019). Guess my power: A computational model to simulate a partner's behavior in the context of collaborative negotiation. In Kohei Arai, et al., editors, *Intelligent Systems and Applications*, pages 1317–1337, Cham. Springer International Publishing.

Piolle, G. and Demazeau, Y. (2008). Une logique pour raisonner sur la protection des donnÃ©es personnelles. In *16e congrÃšs francophone AFRIF-AFIA sur la Reconnaissance de Formes et l'Intelligence Artificielle $RFIA'08$*, page 8p., Amiens, France, Jan. AFRIF - AFIA.

Rajendran, J., Ganhotra, J., Singh, S., and Polymenakos, L. (2018). Learning end-to-end goal-oriented dialog with multiple answers. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3834–3843, Brussels, Belgium, October-November. Association for Computational Linguistics.

Rajendran, J., Ganhotra, J., and Polymenakos, L. (2019). Learning end-to-end goal-oriented dialog with maximal user task success and minimal human agent use. *CoRR*, abs/1907.07638.

Schaub, L.-P. and Vaudapiviz, C. (2019). Goal-oriented dialogue systems : state-of-the-art and future works. In *RECITAL*, Toulouse, France, July.

Serban, I. V., Sordoni, A., Lowe, R., Charlin, L., Pineau, J., Courville, A., and Bengio, Y. (2017). A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, pages 3295–3301. AAAI Press.

Spiekermann, S. (2012). The challenges of privacy by design. *Commun. ACM*, 55(7):38–40, July.

Tahara, S., Ikeda, K., and Hoashi, K. (2019). Empathic dialogue system based on emotions extracted from tweets. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, pages 52–56, New York, NY, USA. ACM.

Wu, C., Madotto, A., Winata, G. I., and Fung, P. (2018). End-to-end dynamic query memory network for entity-value independent task-oriented dialog. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6154–6158, April.

Wu, Y., Wu, W., Xing, C., Xu, C., Li, Z., and Zhou, M. (2019). A sequential matching framework for multi-turn response selection in retrieval-based chatbots. *Computational Linguistics*, 45(1):163–197, March.

Young, S., Gasic, M., Thomson, B., and Williams, J. (2012). Pomdp-based statistical spoken dialogue systems: a review. In *Proceedings of the IEEE*, pages 1–20, January. DOI 10.1109/JPROC.2012.2225812.

Zhang, Z., Huang, M., Zhao, Z., Ji, F., Chen, H., and Zhu, X. (2019). Memory-augmented dialogue management for task-oriented dialogue systems. *ACM Trans. Inf. Syst.*, 37(3):34:1–34:30, July.

*Proceedings of the workshop on Legal and Ethical Issues (Legal2020)*, pages 9–14
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020


# Anonymization for the GDPR in the Context of Citizen and Customer Relationship Management and NLP

**Gil Francopoulo, Léon-Paul Schaub**
Akio + Tagmatica, Akio + LIMSI-CNRS
43 rue de Dunkerque, 75010 Paris, France
gil.francopoulo@wanadoo.fr, lpschaub@akio.com

**Abstract**

The General Data Protection Regulation (GDPR) is the regulation in the European Economic Area (EEA) law on data protection and privacy for all citizens. There is a dilemma between sharing data and their subjects' confidentiality to respect GDPR in the commercial, legal and administrative sectors of activity. Moreover, the case of text data poses an additional difficulty: suppressing the personal information without deteriorating the semantic argumentation expressed in the text in order to apply a subsequent process like a thematic detection, an opinion mining or a chatbot. We listed five functional requirements for an anonymization process but we faced some difficulties to implement a solution that fully meets these requirements. Finally, and following an engineering approach, we propose a practical compromise which currently satisfies our users and could also be applied to other sectors like the medical or financial ones.

**Keywords:** anonymization, pseudonymization, GDPR, NLP

## 1. Introduction

The General Data Protection Regulation (GDPR)[1] is the regulation in the European Economic Area[2] (EEA) law on data protection and privacy for all citizens. The aim is to give control to individuals over their personal data and to simplify the regulatory environment for international business by unifying the regulation within the EEA. The main evolutions brought by the law are:

- New concepts are created or instated: profiling, right to be forgotten, privacy by design (Spiekermann, 2012),
- Transparency becomes an obligation (Goddard, 2017),
- Responsibilities are re-balanced (Lindqvist, 2017).

The regulation contains provisions and requirements of personal data of individuals and applies to any enterprise established in the EEA countries. This regulation changes the way we manage our data (Kamocki et al., 2018)(de Mazancourt et al., 2015). Business processes that handle personal data must be designed with consideration of the principles and provide safeguards to protect data, for example using anonymization, so that the data sets are not publicly available without explicit and informed consent. De-identification like data anonymization is the process of removing personally identifiable information from data sets, so that people whom the data describe remain anonymous (Ji et al., 2017).

Fully anonymized data that meet the legal bar set by European data protection law is no longer 'personal data' and is therefore not subject to the obligations of the GDPR at all. It should be added that a process akin to anonymization is pseudonymization in which personable identifiable information are replaced by one or more artificial surrogates. Pseudonymization[3] is defined in the GDPR Article 4(5) as:

> the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person.

A single pseudonym for each replaced field makes the data record unidentifiable while remaining suitable for data analysis. The main point to consider is the validity coverage of the pseudonym. There are two sorts of pseudonymization: one could be called as 'local' in which the pseudonym is only valid within a single text and the second could be called as 'global' in which the pseudonym is constant from a text to another within a corpus. The main difference between these three options is that anonymous and local pseudonymization data can't be re-identified while global pseudonymization still allows for some re-identification because various clues may be picked and linked together.

Two modes of treatment are concerned:

- During the development phase, in batch mode, large collections of texts need to be collected in order to feed various machine learning processes and statistical computations,
- During exploitation, in real-time mode, a constant flow of information needs to be inserted into real time data analysis or chatbots.

[1] Council Regulation 2016/679, 2016 O.J. (L 119) (EU) 1.

[2] Let's recall that the EEA is the European Union plus Iceland, Liechtenstein and Norway.

[3] Pseudonymization becomes now an active field of research to such an extent that a workshop has just been devoted to it (Ahrenberg and Megyesi, 2019)

The focus of this article can be summed up as: **how to adjust the cursor, in order to respect the personal privacy of the citizens while allowing in depth semantic data analysis at the level of a large group of people and texts?**

## 2. Industrial context

The context is the design and use of an anonymization tool within CRM which means usually Customer Relationship Management for private companies (Garcia-Crespo et al., 2010) but when applied to administration can be formulated as Citizen Relationship Management. The content is either email messages, social media flows or chatbot dialogues.

We operate in both the private and the public domains. In a private context, we work in the domain of e-commerce and retail where NLP techniques are used to compute customer satisfaction (or dissatisfaction) features under GDPR (Sun et al., 2017). In a public context, the communication department of the Prime Minister of an important EEA country receives hundreds of personal complaints and questions per day within a secure perimeter implemented with on-premise servers and firewall protection. The problem arises when these data should be given to another administrative department or to external sub-contractors for data analysis purposes in order to understand what are the concerns of the population directly from the verbatim corpus using NLP techniques. Up until now, this externalization was not possible under GDPR.

## 3. Related works

The problem of customer data anonymization is older than GDPR. (Zhong et al., 2005) show the efficiency of k-anonymization for customer privacy during automatic process. K-anonymization is defined as: 'Given person-specific field-structured data, produce a release of the data with scientific guarantees that the individuals who are the subjects of the data cannot be re-identified while the data remain practically useful' (Samarati and Sweeney, 1998). Although (Nergiz and Clifton, 2007) outperformed k-anonymization with clustering-based algorithms. However these techniques were not effective to anonymize unstructured data as shown by (Angiuli and Waldo, 2016) and the GDPR introduced several changes in the definition of an anonymized text (Hintze and El Emam, 2018).

(Di Cerbo and Trabelsi, 2018) introduce an overview of supervised techniques for anonymization. In the medical domain, NLP tasks are grandly concerned by GDPR (Fraser et al., 2019; Kirinde Gamaarachchige and Inkpen, 2019; Berg and Dalianis, 2019; Dalianis, 2019). (Chevrier et al., 2019) propose a survey on specific techniques and issues of anonymization for medical datasets. (Goddard, 2017) propose a clustering approach for medical reports anonymization in order to limit the information loss and the data utility.

In the didactic field, (Megyesi et al., 2018) build a GDPR-compliant corpus for foreign language learner: their method can be partially reused in many domain because of the complete named-entities anonymization they realize.

Several open-source tools recently appear to anonymize texts according to GDPR (Adams et al., 2019; Kleinberg et al., 2017). Nevertheless, as far as we know, there is no formal approach of text anonymization for opinion mining-based tasks in the customer-relationship management field. Some works processing customer data just could not anonymize their corpora because of the task complexity : (Bechet et al., 2012) developped a corpus of call-center human-human spoken conversation from the Parisian public transport network (RATP) but were not able to distribute it because of the absence of anonymization. Moreover, the GDPR was not voted yet, we guess that anonymization of such a corpus would be even harder today.

## 4. Requirements

The main functional requirements are as follows:

- REQ#1 Avoid identifying the individuals mentioned in the text,
- REQ#2 Allow in-house semantic data analysis which could eventually be adapted to a certain kind of input,
- REQ#3 Allow off-the-shelf NLP tools,
- REQ#4 Prove that an anonymization has been done in case of a complaint from someone mentioned in a specific text or in case of lawsuit or journalists' investigation,
- REQ#5 Usable in different European languages.

These requirements are somehow contradictory. For instance, from the original text:

> My name is Paul Smith, and I moved from Leeds to Paris.

an anonymization will black out all identifiable information and will produce:

> My name is X and I moved from X to X.

In this case, REQ#1 and REQ#4 are fulfilled but the semantic processing of REQ#2 and REQ#3 will be deeply disrupted. Another option could be to replace a name with a random name from a dictionary while respecting the type of the name as:

> My name is John Wilson and I moved from Berlin to Madrid.

In this case, the realistic surrogates give the impression that the text is original but REQ#4 is not fulfilled. We cannot afford global pseudonymization because it is not really a secure anonymization (as mentioned in the introduction) but local pseudonymization seems a good compromise fulfilling four out of five requirements giving a sentence like:

> My name is _People1 and I moved from _City1 to _City2.

| Req. | substitution by X | global pseudo. | local pseudo. | random substitution |
|---|---|---|---|---|
| REQ#1 | yes | no | yes | yes |
| REQ#2 | no | yes | yes | yes |
| REQ#3 | no | no | no | yes |
| REQ#4 | yes | yes | yes | no |
| REQ#5 | yes | yes | yes | yes |

Table 1: Requirements vs solutions

The only drawback of the approach is that the text cannot be given to an NLP process which is not prepared for this sort of mangling like an automatic translation, and therefore the REQ#3 target is missed. In fact, **REQ#3 and REQ#4 are contradictory**. Thinking again about this problem, we realized that certain requirements need not to be satisfied in all circumstances. REQ#4 is important when producing the data out of a secure perimeter while REQ#3 is important when using off-the-shelf tools internally within a secure perimeter. The dilemma can be resolved by implementing a Boolean parameter when running the anonymization associated to REQ#3 or REQ#4 fulfillment. Thus, the anonymization is able to produce:

> My name is _People1 and I moved from _City1 to _City2.

when there is a need to externalize, as well as:

> My name is John Wilson and I moved from Berlin to Madrid.

in case of internal processing, depending on the option. The requirement fulfillment is summed up in table-1.

## 5. Implementation

The idea is to chain three processes: 1) a named entity recognition, 2) an entity linker, and 3) a substitution. These processes should run within a secure environment and should not produce any traces of execution which could break the anonymization. That is to say that only the result of the substitution is authorized to be publish outside the running environment.

Named entity recognition (NER) is processed by Akio's named entity detector which takes the output of a syntactic parser whose name is Tagparser (Francopoulo, 2008). The parser combines statistical induction and robust syntactic rules. The NER is implemented by a cascade of patter matching rules to detect names of human beings, locations, companies, marks, email addresses and all sorts of numeral forms like dates, amounts of money, flight numbers, IBANs, phone numbers, passport numbers and social security numbers[4]. For proper names, the NER makes use of language-based local clues combined with a list of 1.2M proper names which have been automatically extracted from Wikidata. This is an industrial detector used to process currently an average of 1M texts every day in six languages (English, French, German, Italian, Spanish, Portuguese). There is a specific parser for each language whereas most named entity detections are language-neutral, that is there are the same in all our six covered languages. In fact, only a small set of cultural differences like vehicle identifications are different[5]. The program includes a specific spelling checker to process ill-formatted inputs based on a 10 years' experience of badly formatted input collection.

The aim of the entity linker is to gather named entities appearing in different places of the text possibly with some encyclopedic or orthographic variations. For instance, in 'Nicolas Sarkozy said…Sarko replied…' where 'Sarko' being a nickname for 'Nicolas Sarkozy', the two names should be linked. Another example is 'N Sarkozy' vs 'Nicolas Sarkozy' where 'N' is not ambiguous and should be considered as a given name. The objective is to link these utterances in a common structure.

The objective of the substitution is to replace a selection of entity types which are:

- **city** for the names of cities and agglomerations, like 'Paris' (a city) or 'Cergy-Pontoise' which is not formally a city but is an agglomeration.
- **contractNumber** for the combination of digit and letters which seems to be something else than a word or a number. This category includes some specific personal categories like IBANs (International Bank Account Number) and BICs (Bank Identifier Code).
- **emailAddress** for email addresses.
- **personName** for the names of individuals which are human beings.
- **identificationNumber** for the identifier of an individual like a social security number or a passport number.
- **IPAddress** for Internet Protocol addresses.
- **phoneNumber** for the various forms of a phone number.
- **vehicleIdentification** for the vehicle registration plates.
- **zipCode** for postal codes.

It should be noted that the NER detects other entity types like for instance, countries, regions, organizations, amounts of money or flight numbers. Obviously, it is technically easy to substitute these entities but the question is: what is the rationale to do so? These entities are less personal and without any personal clues there is no danger in keeping

[4]The reader can reproduce our work by using another NER provided that all the precise and personal forms like social security identifiers are correctly detected. Obviously, the quality of the whole process is highly dependent on that of the NER.

[5]The French system is not able to recognize German number plates, for instance, but the situations where it is necessary are extremely rare.

the original string, provided that the more the text is transformed, the more difficult the semantic parsing is. Due to the fact that city is replaced, the exact localisation cannot be determined, so there is no need to substitute the address in full, in addition to the fact that the recognition of the section indicating the street is very difficult because of the many possible forms.

## 6. Example

From this (invented) original text:

> Dear Sir/Madam,
> I am writing today to complain of the problem I have with www.ameli.fr. I'd like to create an account but my social security identifier 200 11 99 109794 on my carte vitale is not the same as the one of my mutual insurance 201 11 99 109794. How could I do?
> Best regards,
> Paul Watson,
> tel 01 23 34 34 56 pwatson@aol.fr

Note that the Carte Vitale is the health insurance card of the national health care system in France. The anonymization produces the following text, provided that the pseudonymization option is selected:

> Dear Sir/Madam,
> I am writing today to complain of the problem I have with www.ameli.fr. I'd like to create an account but my social security identifier _SSid1 on my carte vitale is not the same as the one of my mutual insurance _SSid2 How could I do?
> Best regards,
> _People1,
> tel _Phone1 _Email1

Due to the fact that the pseudonyms are renumbered starting at one in each text, it is not possible to induce any personal data from this text or to make any correlation with another text, so GDPR is respected. However, provided that the digital analytics program is specially adapted to orthographically handle pseudonyms and to interpret the pseudonym as a semantic named entity value, it is still possible to compute that the author has:

- A complaint concerning a given web site,
- A complaint of mismatch concerning different social security identifiers,
- A question.

This is fully satisfactory. It is typically the kind of results which are produced by our in-house product Akio Analytics but such a result could also be computed by another product implementing ABSA (Aspect Based Sentiment Analysis) as we do (Pontiki et al., 2014).

## 7. Method used for validation

The manual verification of a large corpus iteratively with alternations of correction / verification is a very heavy burden. Our test corpus is a collection of 18138 French verbatim from the legal and administrative sector of activity and we cannot verify the whole corpus after every improvement of the detector. We started by excluding randomly 300 verbatim as a test corpus, to be used afterwards.
The main focus is not to avoid noise but mainly to avoid silence, that is, we consider that it is not very important when a character string is over-substituted. On the contrary, missing a person name substitution is a serious mistake. We use the fact that the text is transformed after pseudonymization and if some proper names of the nine types are remaining, there is a good chance that there is an error. We tested the system for French following a three-fold approach iteratively on the development corpus containing 18138-300=17838 texts:

- Step#1, the corpus is anonymized with the local pseudonymization option,
- Step#2 the named entity is applied again and the result is filtered to retain the named entities of the nine types which do not begin with the character underscore, this character identifying a pseudonym. When there is a result, there is a good chance that this is an error.
- Step#3 the NER errors are fixed and the process is applied again at Step#1. We stopped when we have not found any error.

The different phases of the validation are presented in table-2.

| rounds | nb of processed texts | nb of errors |
|---|---|---|
| phase-1 | 17838 | 284 |
| phase-2 | 284 | 53 |
| phase-3 | 53 | 0 |

Table 2: Results of validation

Evaluation of the test corpus is presented in table-3:

| Nb of texts | Recall | Precision | FMesure |
|---|---|---|---|
| 300 | 100 | 99.5 | 99.7 |

Table 3: Quality evaluation

The total distribution over the whole corpus (development and test) by type of entity is shown in Table 4.

## 8. Future work

The NER is currently used everyday in order to compute e-reputation and commercial data analysis in six languages for several big companies, but so far, we did not had time to work on anonymization in all these languages. In the near future, we plan to test the anonymization in languages other than French.
We also plan to extend the substitution to another entity which does not directly identify an individual but which by its context can do so, what is usually called a context-sensitive entity. We plan to substitute all organizational

| type | nb of occ. | distrib. |
|---|---|---|
| city | 6408 | 19% |
| contractNumber | 17 | 0% |
| emailAddress | 2141 | 6% |
| personName | 20835 | 63% |
| identificationNumber | 146 | 0% |
| IPAddress | 89 | 0% |
| phoneNumber | 1721 | 5% |
| vehicleIdentification | 97 | 0% |
| zipCode | 1687 | 5% |
| total | 33141 | 100% |

Table 4: Entity types distribution

names for (relatively rare) cases like: "as president of Danone".

## 9. Conclusion

After a presentation of the context of use which is rather broad, namely citizen and customer relationship management, we listed five precise requirements and discussed the various options to provide an effective implementation. Our requirements are not specific to our context and could be applied to another context like a medical or financial application.

Our process anonymizes critical information through a step-wise named entity recognition implementation and entity linking. It identifies contextual information and replaces them with a semantic-preserving category label which allow semantic data analytics except that the character string of certain proper names and numeric expressions are hidden but remain manageable. As an option, the program allows the replacement with a random value simulating an original character string for off-the-shelf NLP tools.

## 10. Acknowledgements

This work was co-financed by ANRT and Akio under the CIFRE contract 2017/1543.

## 11. Bibliographical References

Adams, A., Aili, E., Aioanei, D., Jonsson, R., Mickelsson, L., Mikmekova, D., Roberts, F., Valencia, J. F., and Wechsler, R. (2019). AnonyMate: A toolkit for anonymizing unstructured chat data. In *Proceedings of the Workshop on NLP and Pseudonymisation*, pages 1–7, Turku, Finland, 30 September. Linköping Electronic Press.

Lars Ahrenberg et al., editors. (2019). *Proceedings of the Workshop on NLP and Pseudonymisation*, Turku, Finland, 30 September. Linköping Electronic Press.

Angiuli, O. and Waldo, J. (2016). Statistical tradeoffs between generalization and suppression in the deidentification of large-scale data sets. In *2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC)*, volume 2, pages 589–593, June.

Bechet, F., Maza, B., Bigouroux, N., Bazillon, T., ElbÃ¨ze, M., Mori, R. D., and Arbillot, E. (2012). Decoda: a call-center human-human spoken conversation corpus. In *In International Conference on Language Resources and Evaluation (LREC*.

Berg, H. and Dalianis, H. (2019). Augmenting a deidentification system for Swedish clinical text using open resources and deep learning. In *Proceedings of the Workshop on NLP and Pseudonymisation*, pages 8–15, Turku, Finland, 30 September. Linköping Electronic Press.

Chevrier, R., Foufi, V., Gaudet-Blavignac, C., Robert, A., and Lovis, C. (2019). Use and understanding of anonymization and de-identification in the biomedical literature: Scoping review. *J Med Internet Res*, 21(5):e13484, May.

Dalianis, H. (2019). Pseudonymisation of Swedish electronic patient records using a rule-based approach. In *Proceedings of the Workshop on NLP and Pseudonymisation*, pages 16–23, Turku, Finland, 30 September. Linköping Electronic Press.

de Mazancourt, H., Couillault, A., Adda, G., and Recourcé, G. (2015). Faire du TAL sur des données personnelles : un oxymore ? In *22eme Conférence sur le Traitement Automatique des Langues Naturelles*, Caen, France, June.

Di Cerbo, F. and Trabelsi, S. (2018). Towards personal data identification and anonymization using machine learning techniques. In András Benczúr, et al., editors, *New Trends in Databases and Information Systems*, pages 118–126, Cham. Springer International Publishing.

Francopoulo, G. (2008). Tagparser: well on the way to ISO-TC37 conformance. In *ICGL (International Conference on Global Interoperabilty for Language Resources)*, Hong Kong, January.

Fraser, K. C., Linz, N., Lindsay, H., and König, A. (2019). The importance of sharing patient-generated clinical speech and language data. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 55–61, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Garcia-Crespo, A., Colomo-Palacios, R., Gomez-Berbis, J. M., and Ruiz-Mezcua, B. (2010). Semo: A framework for customer social networks analysis based on semantics. *Journal of Information Technology*, 25(2):178–188.

Goddard, M. (2017). The EU general data protection regulation (GDPR): European regulation that has a global impact. *International Journal of Market Research*, 59(6):703–705.

Hintze, M. and El Emam, K. (2018). Comparing the benefits of pseudonymisation and anonymisation under the GDPR. *Journal of Data Protection & Privacy*, 2(2):145–158.

Ji, S., Mittal, P., and Beyah, R. (2017). Graph data anonymization, de-anonymization attacks, and de-anonymizability quantification: A survey. *IEEE Communications Surveys Tutorials*, 19(2):1305–1326, Secondquarter.

Kamocki, P., Mapelli, V., and Choukri, K. (2018). Data management plan (DMP) for language data under the new general da-ta protection regulation (GDPR). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Kirinde Gamaarachchige, P. and Inkpen, D. (2019). Multi-task, multi-channel, multi-input learning for mental illness detection using social media text. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 54–64, Hong Kong, November. Association for Computational Linguistics.

Kleinberg, B., Mozes, M., Toolen, Y., and Verschuere, B. (2017). Netanos - named entity-based text anonymization for open science. June.

Lindqvist, J. (2017). New challenges to personal data processing agreements: is the GDPR fit to deal with contract, accountability and liability in a world of the Internet of Things? *International Journal of Law and Information Technology*, 26(1):45–63, December.

Megyesi, B., Granstedt, L., Johansson, S., Prentice, J., Rosén, D., Schenström, C.-J., Sundberg, G., Wirén, M., and Volodina, E. (2018). Learner corpus anonymization in the age of GDPR: Insights from the creation of a learner corpus of Swedish. In *Proceedings of the 7th workshop on NLP for Computer Assisted Language Learning*, pages 47–56, Stockholm, Sweden, November. LiU Electronic Press.

Nergiz, M. E. and Clifton, C. (2007). Thoughts on k-anonymization. *Data & Knowledge Engineering*, 63(3):622 – 645. 25th International Conference on Conceptual Modeling (ER 2006).

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., and Manandhar, S. (2014). SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland, August. Association for Computational Linguistics.

Samarati, P. and Sweeney, L. (1998). Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report.

Spiekermann, S. (2012). The challenges of privacy by design. *Commun. ACM*, 55(7):38–40, July.

Sun, S., Luo, C., and Chen, J. (2017). A review of natural language processing techniques for opinion mining systems. *Information Fusion*, 36:10 – 25.

Zhong, S., Yang, Z., and Wright, R. N. (2005). Privacy-enhancing k-anonymization of customer data. In *Proceedings of the Twenty-fourth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '05, pages 139–147, New York, NY, USA. ACM.

*Proceedings of the workshop on Legal and Ethical Issues (Legal2020)*, pages 15–20
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020


# Ethically Collecting Multi-Modal Spontaneous Conversations with People that have Cognitive Impairments

**Angus Addlesee, Pierre Albert**
Heriot-Watt University, University of Edinburgh
Edinburgh, UK, Edinburgh UK
ja204@hw.ac.uk, pierre.albert@ed.ac.uk

**Abstract**

In order to make spoken dialogue systems (such as Amazon Alexa or Google Assistant) more accessible and naturally interactive for people with cognitive impairments, appropriate data must be obtainable. Recordings of multi-modal spontaneous conversations with vulnerable user groups are scarce however and this valuable data is challenging to collect. Researchers that call for this data are commonly inexperienced in ethical and legal issues around working with vulnerable participants. Additionally, standard recording equipment is insecure and should not be used to capture sensitive data. We spent a year consulting experts on how to ethically capture and share recordings of multi-modal spontaneous conversations with vulnerable user groups. In this paper we provide guidance, collated from these experts, on how to ethically collect such data and we present a new system - "CUSCO" - to capture, transport and exchange sensitive data securely. This framework is intended to be easily followed and implemented to encourage further publications of similar corpora. Using this guide and secure recording system, researchers can review and refine their ethical measures.

**Keywords:** ethical data collection, multi-modal interaction, data security

## 1. Introduction

In this paper, we first introduce the background and motivations behind our work before detailing our contributions to ethical protocols in section 2. We provide guidance, collated from meetings with experts, on ethically collecting multi-modal spontaneous conversations with people that have cognitive impairments. We have also created a system to securely record this data. This new system is detailed in section 2.5.

### 1.1. Dialogue as Cognition Declines

Natural face-to-face conversations involve quick exchanges that are littered with hesitations, restarts, self-corrections (Shriberg, 1996; Hough, 2015), interruptions (Healey et al., 2011), backchannels (Heldner et al., 2013; Howes and Eshghi, 2017) and split utterances (Howes, 2012), etc... with none of these phenomena respecting the boundaries of a sentence or turn. These phenomena become even more common and more pronounced as cognition declines. For example, people with certain types of dementia pause more frequently and for longer durations than healthy controls (Boschi et al., 2017). These changes have even been used successfully to detect Alzheimer's disease (AD) from just a person's speech (Luz et al., 2018; Zhu et al., 2018).

People don't just communicate using words however, visual feedback (nodding, brow furrowing, head tilting, etc...) and hesitation utterances ("umm", "err", "hmm", etc...) are short but do guide conversation (Goodwin, 1981; Bavelas and Gerwing, 2011). Whether non-verbal interactions change as cognition declines is relatively unknown because multi-modal recordings of such interactions are scarce.

### 1.2. Older Adults & Spoken Dialogue Systems

When people speak to Spoken Dialogue Systems (SDSs), such as Amazon Alexa or Google Assistant, they adapt to the system (Pelikan and Broth, 2016; Porcheron et al., 2018). Each utterance is stripped of the phenomena discussed in section 1.1. Adapting to an SDS is acceptable for the majority of users but older adults, with less exposure to such systems, and people with cognitive impairments can struggle to adapt their natural interaction patterns.

The global ageing population (UN, 2018) relies on care systems that have been strained for years (Wright, 2015). This causes knock-on problems, such as bed-blocking in hospitals (Puttick, 2018), and these pressures can be eased if people are able to live in their own homes for longer and more independently. A huge range of IoT devices could help tackle this challenge but their embedded SDSs need to become more natural if they are to make an impact (Sakakibara et al., 2017; Helal and Bull, 2019).

### 1.3. Our Corpus Collection Details

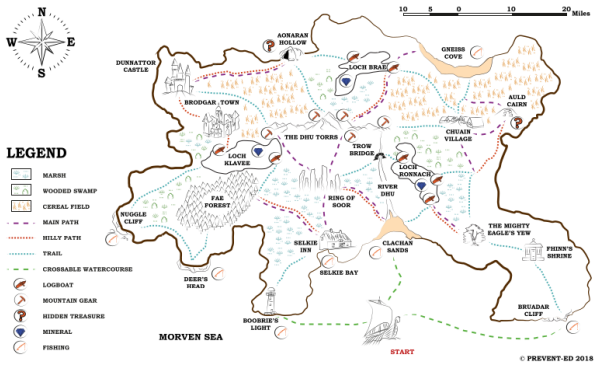


Figure 1: PREVENT-Elicitation of Dialogues (PREVENT-ED) map with routes (de la Fuente Garcia et al., 2019).

For context, we are collecting a corpus of conversations with people that have various types of dementia. This does not constrain our work however and the ethical protocol extends to cover cognitive impairments more generally, estab-

lished in Section 2.

Dementia is one of the leading causes of death in the UK (Alzheimer's Research UK, 2018) but there is no treatment to prevent, cure or slow its progression. Even with suitable data, adapting SDSs for those living with dementia is not trivial due to the many challenges left to tackle (Addlesee et al., 2019). Monologue recordings of people with AD describing pictures exist (Becker et al., 1994) and interviews of people with AD exist (Pope and Davis, 2011) which are both used to develop the AD detection models mentioned in section 1.1. Neither of these corpora contain the spontaneous conversational speech that we would expect an SDS to receive and, importantly, they are only audio recordings. Therefore, we are going to collect a multi-modal corpus of spontaneous conversations with people that have various types of dementia.

A variant of the map task (Anderson et al., 1991) has been recently developed to elicit spontaneous spatial navigation dialogues with people that have dementia (de la Fuente Garcia et al., 2019) and we are collaborating with the creators of this task to collect our corpus. Using this task, a healthy participant will sit opposite a person that has dementia for a casual conversation. Both participants have a map with the same locations, but only the person with dementia can see the possible routes through the imaginary land (as shown in Figure 1). The healthy participant, however, is the only one who knows which locations the pair need to visit. They therefore need to collaborate through conversation to go on the journey together.

## 2. Ethical Considerations

Over the past year, we have contacted and met with many experts to ensure that we collect our corpus ethically. These experts belong to many institutions including: The NHS, Alzheimer Scotland, Edinburgh Medical School, Edinburgh Centre for Dementia Prevention, Heriot-Watt University and more. We have collated all the information in this paper and we have also developed a new system to record multi-modal interactions more securely. This system is detailed in section 2.5.

### 2.1. Consent

Each participant will be given a participant information sheet (PIS) before taking part in the study. This PIS contains all information about the study (what it will involve, the benefits of taking part, what data will be stored, etc...) and should be given to the participant at least a week before they take part. This time allows the participant to digest the information and ask any questions to family members, carers, GPs, or a member of the research team. A consent form is then provided before the experiment that summarises the key points in the PIS and confirms that the participant has read and understood it. It is important to stress that all questions are welcome and that participation is entirely voluntary. The documents are distributed as shown in Figure 2.

People with cognitive impairments are considered vulnerable participants and a witness is therefore required. The witness should be a family member or carer (Lacny et al., 2012) and has to sign a witness of consent form. This should be signed *after* the participant signs their consent

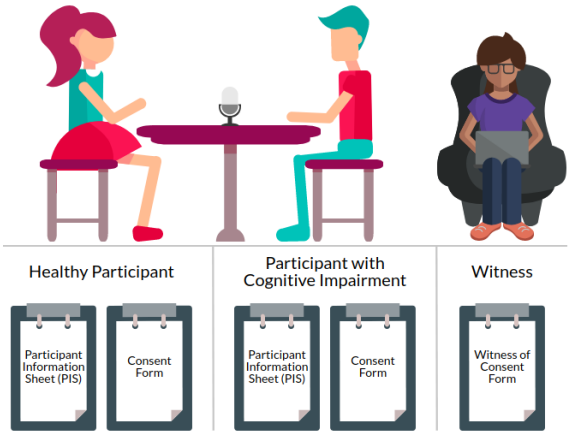


Figure 2: The distribution of required documentation.

form as it confirms that they understood the PIS, had all of their questions answered, and willingly consented to take part in the study. Immoral researchers could attempt to trick a person with a cognitive impairment (for example, offering to make them a cup of tea after they "sign a quick form") or elicit personal information (for example, asking about their previous medical history). To ensure this cannot happen, the witness also signs to confirm that the researcher did not attempt to elicit personal information, mislead, or trick the participant.

### 2.2. Participant Comfort

Participants are spending their valuable time helping with research but could feel stressed about taking part, especially those with cognitive impairments. Ensuring people have a comfortable experience is therefore of paramount importance.

Even before taking part, the PIS should contain as much information as possible to prevent unnecessary stress. For example, it can highlight the following about the task:

- It requires no preparation.
- It is not a medical examination.
- We want to record a natural conversation, so it is intended to be a fun game.
- Recording can be stopped (or paused) at any time without giving a reason.
- There is no right or wrong answer.
- There is no time limit.

Some people may feel uncomfortable stopping the study, even if they are feeling distressed. A family member or carer should witness the task for this reason, usually the same witness that we discussed in section 2.1. The witness can also stop or pause the recording at any point without giving a reason. As a researcher, it is crucial to understand the importance of this witness. Different cognitive impairments and even different people with the same cognitive impairment have distinct signals to indicate distress. Family members and carers are significantly more experienced

at identifying whether a particular person is uncomfortable, than any researcher, because they know that exact person.

Figure 3: An Alzheimer Scotland café, designed to be an accessible community hub (Graven, 2014).

A suitable location is relatively easy to find as spontaneous conversations can take place almost anywhere. For participant comfort and availability of a witness however, it is best to collaborate with a business or charity focused on the cognitive impairment of interest. To engage with their communities, these organisations usually have drop-in centres that people can visit for social activities, support, and classes (an example is shown in Figure 3). These centres are perfect locations to run tasks as people are very comfortable in them. The staff also know the potential participants and can therefore be witnesses and help with recruitment, discussed in section 2.3. Most accessible locations are suitable but working with a charity, to carry out the study in one of their centres, is the best option when possible.

### 2.3. Participant Recruitment

Collaborating with a relevant organisation is vital when recruiting vulnerable participants, this is in addition to the benefits around participant comfort. These organisations can reach out to their community and assist with recruitment of suitable participants in a safe and friendly manner. Collaboration costs the organisation valuable time however, so it is important to explain the motivations behind the data collection. We have had very positive responses from multiple charities using the rationale given in section 1.

Healthy participants are also required to partake as interlocutors. It is common to compensate research participants with small rewards, such as gift cards, but it is not advised in this case. People with cognitive impairments will be using their time to contribute to research by taking part in the task. The healthy participant will ideally be motivated by the contribution to society and not some end reward. Someone who does not care about the motivations behind the research could rush through the task for a gift card, devaluing the vulnerable participants time. This example case can not happen if there is no monetary reward offered for taking part.

### 2.4. Optional Cognitive Assessment

Collecting multi-modal recordings of these conversations is a long and costly process. It is therefore worthwhile to share this data; and we will do so as detailed in section 2.6.. For use by other researchers in certain fields, such as Psychology, cognitive assessment results have huge benefits. For example, another corpus that performed the same cognitive assessment could be merged to reveal unknown connections. There are also worries to consider before including such a task however.

For example, the task that is most suitable for our data collection is the Addenbrooke's Cognitive Examination (ACE-III) (Hsieh et al., 2013) as it is commonly used, low-tech, and quick to perform. An example question can be seen in Figure 4. Importantly, NHS training needs to be passed in order to run this test and it should not be recorded audiovisually. Similarly for other tests, all training must be completed prior to collection. One downside to highlight is that the ACE-III is used by GPs to screen people while diagnosing dementia. Therefore, participants may recall doing this task and be reminded of the stressful times around their diagnosis. This could upset a participant and in addition, retaking the test may highlight how they have declined in cognitive performance since first completing it.

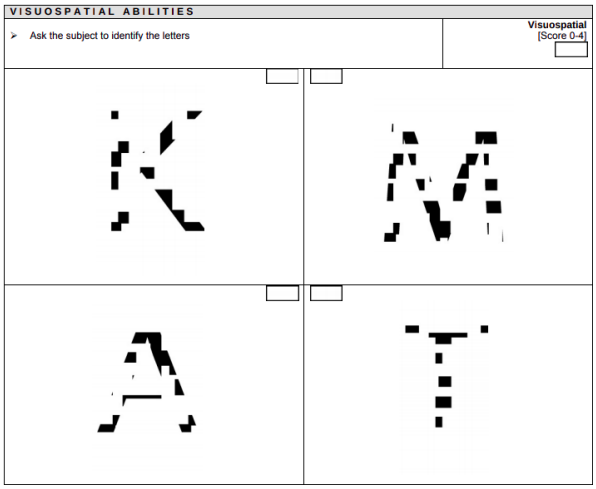


Figure 4: An example question from the ACE-III.

Each cognitive impairment has a range of tests to scrutinise and it is very valuable to include a cognitive assessment. A person's well-being should be prioritised however, so only run tests after careful consideration and the relevant training.

### 2.5. Securely Recording Multi-Modal Interactions

Conversations involving patients or medical personnel are full of sensitive data. People often disclose personally identifiable information during a conversation (for example, mentioning their children's names or medical history). This concern is even stronger with conversations involving vulnerable participants (e.g. people with cognitive impairments), less prone to control the information they disclose. Standard recording systems (e.g. audio recorders and video

cameras) are not secure devices, and they cannot be used to capture sensitive data. Furthermore, recorded data can easily be accessed on standard systems. Ethical and legal consequences of data breach must be accounted for if a standard device is lost or stolen, highlighting the need for a secure approach.

A new system - "CUSCO" - was developed to satisfy the requirements stemming from the ethical assessment regarding data collection of sensitive material. The device allows the collection of a range of modalities, including audio and video. Handling conversations containing sensitive material requires mitigation of the consequences of unintentional or fraudulent loss of data. The device ensures the security of the recorded data by encrypting recorded streams in real time. The encryption is done using Veracrypt, a dedicated open-source software that underwent a security audit, vouching for the correct implementation of the encryption algorithms.

Collected data can only be accessed with the key generated for each project, ensuring security of the corpus during *all* the phases of its life: collection, transport, exchange and storage.

The CUSCO device was designed to collect medical conversations between healthcare professionals and patients *in-situ*. Recording material in medical practices is common to study real-life phenomena (Montague and Asan, 2014), but - to our knowledge - considerations for the security of the collected data are overwhelmingly ignored.

Recent legal evolution on the protection of personal data, such as the General Data Protection Regulation (GDPR - (Parliament and the Council, 2016)) in the EU, has led to a strong focus being put on these considerations during ethical evaluation and validation of new research projects involving data collection, use, and sharing.

Risk prevention and mitigation for data handling set the functional requirements for the design of our system. As such, even if the device is compromised or stolen *during* recording, the entire dataset of previously recorded conversations and any recording in progress are secure.

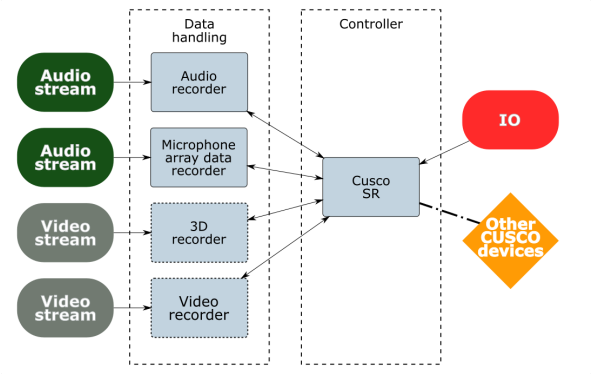


Figure 5: Main components of the CUSCO device.

Furthermore, data collections can have stricter ethical controls: researchers may also be to prohibited direct access to sensitive information. The device provides capabilities for the collection of anonymised audio and visual features, i.e. an abstracted indirect description of the interaction that cannot be used to reconstruct the original signal.

The software of the device itself is organised around a modular design, as described in Figure 5. Each stream, corresponding to a modality (video, audio, 3D) or a function (Voice Activity Detection) is controlled by a dedicated module in charge of setting the configuration, checking the state of required elements (presence of the appropriate device), and managing the recording.

For the use-case described in this paper (depicted in Figure 6), we are using two depth cameras, a high-quality table microphone, and a microphone array to facilitate speaker diarisation in post-processing (segmentation of the audio and attribution to each speaker). The decision to use a table microphone was taken because lapel microphones need to be attached to the participants, which can be invasive and can cause distress. The necessity to record high-quality data (use of an additional microphone and recording close-up video of both participants) lead us to reach the capacity of the system and therefore set multiple devices in a network, two in our case.

Figure 6: Setup of the recording system.

The hardware of the device uses common off-the-shelf elements, while the software is open-source. Design schemes and software have been made available online[1]. The need for such a device extends beyond the conversations that we detail in this paper to any sensitive recordings that should be encrypted live. Such use cases include recordings of: GP consultations, interactions with children, and discussions with private companies.

## 2.6. Data Handling and Sharing

Once the conversations have been recorded securely, they remain encrypted on the system detailed in section 2.5. The research team then need to remove any personal information that may have been disclosed during the conversations. To do this, the audio is silenced and the video blurred around the mouth whenever sensitive information is uttered. Blurring video reduces the accuracy of visual behaviour annotation (Lasecki et al., 2015) but privacy takes precedence to avoid possible participant identification. The

[1] https://cybermat.tardis.ed.ac.uk/pial/inca

transcription can therefore not contain any sensitive information (and should not be transcribed from the original recordings to ensure this).

Personal information will not be shared and it is important to highlight this in the PIS discussed in section 2.1. These processed recordings are now considered anonymous as the participants are only identifiable by personal contacts (thus, an unknown researcher cannot identify the participant). The contact details of a member in the research team should be included in the PIS to allow the request for deletion, and subsequent removal, of a participant's data.

The anonymised recordings and associated transcriptions can be shared with other relevant researchers through centralised archives to control its use (Derry et al., 2010), *if* stated in the PIS, and results published in research papers. We have decided to store our corpus in DementiaBank (Becker et al., 1994) as it is a shared database of multimedia interactions for the study of communication in dementia. Access to the data in DementiaBank is password protected and restricted to members of the DementiaBank consortium group. Researchers that would benefit from access to this data can request to join this group and therefore benefit from the corpus.

## 3. Conclusion

Collecting multi-modal spontaneous conversations from people with cognitive impairments is a vital step towards creating more accessible and natural SDSs. To ensure this is done ethically, there are many factors that need to be considered which we have collated and detailed throughout Section 2. This practical ethical framework can assist researchers who want to navigate the many ethical challenges in order to collect and release corpora of multi-modal interactions. Additionally, CUSCO can be used to securely capture, transport and exchange this data.

## 4. Bibliographical References


Addlesee, A., Eshghi, A., and Konstas, I. (2019). Current challenges in spoken dialogue systems and why they are critical for those living with dementia. *Dialog for Good*.

Alzheimer's Research UK, D. S. H. (2018). Deaths due to dementia. *Alzheimer's Research UK*.

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., et al. (1991). The hcrc map task corpus. *Language and speech*, 34(4):351–366.

Bavelas, J. B. and Gerwing, J. (2011). The listener as addressee in face-to-face dialogue. *International Journal of Listening*, 25(3):178–198.

Becker, J. T., Boiler, F., Lopez, O. L., Saxton, J., and McGonigle, K. L. (1994). The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of Neurology*, 51(6):585–594.

Boschi, V., Catricala, E., Consonni, M., Chesi, C., Moro, A., and Cappa, S. F. (2017). Connected speech in neurodegenerative language disorders: a review. *Frontiers in psychology*, 8:269.

de la Fuente Garcia, S., Ritchie, C. W., and Luz, S. (2019). Protocol for a conversation-based analysis study: Prevent-ed investigates dialogue features that may help predict dementia onset in later life. *BMJ open*, 9(3):e026254.

Derry, S. J., Pea, R. D., Barron, B., Engle, R. A., Erickson, F., Goldman, R., Hall, R., Koschmann, T., Lemke, J. L., Sherin, M. G., et al. (2010). Conducting video research in the learning sciences: Guidance on selection, analysis, technology, and ethics. *The Journal of the Learning Sciences*, 19(1):3–53.

Goodwin, C. (1981). *Conversational organization: interaction between speakers and hearers*. Academic Press.

Graven. (2014). Kilmarnock dementia resource centre.

Healey, P. G. T., Eshghi, A., Howes, C., and Purver, M. (2011). Making a contribution: Processing clarification requests in dialogue. In *Proceedings of the 21st Annual Meeting of the Society for Text and Discourse*, Poitiers, July.

Helal, S. and Bull, C. N. (2019). From smart homes to smart-ready homes and communities. *Dementia and geriatric cognitive disorders*, 47(3):157–163.

Heldner, M., Hjalmarsson, A., and Edlund, J. (2013). Backchannel relevance spaces. In *Nordic Prosody: Proceedings of XIth Conference, Tartu 2012*, pages 137–146.

Hough, J. (2015). *Modelling Incremental Self-Repair Processing in Dialogue*. Ph.D. thesis, Queen Mary University of London.

Howes, C. and Eshghi, A. (2017). Feedback relevance spaces: The organisation of increments in conversation. In *IWCS 2017—12th International Conference on Computational Semantics—Short papers*.

Howes, C. (2012). *Coordination in dialogue: Using compound contributions to join a party*. Ph.D. thesis, Queen Mary University of London.

Hsieh, S., Schubert, S., Hoon, C., Mioshi, E., and Hodges, J. R. (2013). Validation of the addenbrooke's cognitive examination iii in frontotemporal dementia and alzheimer's disease. *Dementia and geriatric cognitive disorders*, 36(3-4):242–250.

Lacny, C., Kirk, A., Morgan, D. G., and Karunanayake, C. (2012). Predictors of cognitive impairment severity in rural patients at a memory clinic. *Canadian journal of neurological sciences*, 39(6):774–781.

Lasecki, W. S., Gordon, M., Leung, W., Lim, E., Bigham, J. P., and Dow, S. P. (2015). Exploring privacy and accuracy trade-offs in crowdsourced behavioral video coding. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1945–1954.

Luz, S., de la Fuente, S., and Albert, P. (2018). A method for analysis of patient speech in dialogue for dementia detection. *arXiv preprint arXiv:1811.09919*.

Montague, E. and Asan, O. (2014). Dynamic modeling of patient and physician eye gaze to understand the effects of electronic health records on doctor–patient communication and attention. *International Journal of Medical Informatics*, 83(3):225–234, Mar.

Parliament, T. E. and the Council. (2016). Regulation (eu) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free

movement of such data, and repealing directive 95/46/ec (general data protection regulation). Apr.

Pelikan, H. R. and Broth, M. (2016). Why that nao? how humans adapt to a conventional humanoid robot in taking turns-at-talk. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4921–4932.

Pope, C. and Davis, B. H. (2011). Finding a balance: The carolinas conversation collection. *Corpus Linguistics and Linguistic Theory*, 7(1):143–161.

Porcheron, M., Fischer, J. E., Reeves, S., and Sharples, S. (2018). Voice interfaces in everyday life. In *Proceedings of CHI*.

Puttick, H. (2018). Bed-blocking at worst level since 2016. *The Times*.

Sakakibara, S., Saiki, S., Nakamura, M., and Yasuda, K. (2017). Generating personalized dialogue towards daily counseling system for home dementia care. In *International Conference on Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management*, pages 161–172. Springer.

Shriberg, E. (1996). Disfluencies in switchboard. In *In Proceedings of the International Conference on Spoken Language Processing*, volume 96, pages 3–6. Citeseer.

UN. (2018). World population ageing highlights. *United Nations Department of Economic and Social Affairs*.

Wright, C. (2015). Facts about carers. *Carers UK*.

Zhu, Z., Novikova, J., and Rudzicz, F. (2018). Detecting cognitive impairments by agreeing on interpretations of linguistic features. *arXiv preprint arXiv:1808.06570*.

# Author Index