The Mind-It Corpus: a Longitudinal Corpus of Electronic Messages Written by Older Adults with Incipient Alzheimer's Disease and Clinically Normal Volunteers

Olga Seminck, Louise-Amélie Cougnon, Bernard Hanseeuw, Cédrick Fairon

The Mind-It Corpus: a Longitudinal Corpus of Electronic Messages Written by Older Adults with Incipient Alzheimer's Disease and Clinically Normal Volunteers

Olga Seminck, Louise-Amélie Cougnon, Bernard Hanseeuw, Cédrick Fairon
Université catholique de Louvain
Place de l'Université 1, 1348 Louvain-la-Neuve, Belgium
{olga.seminck, louise-amelie.cougnon, bernard.hanseeuw, cedrick.fairon}@uclouvain.be

Abstract

In this article, we present the Mind-It project and the corpus we are currently collecting. The long-term aim of the project is to contribute to the preclinical detection of Alzheimer's disease (AD) by developing a computer model that searches for linguistic changes that mark AD. To this end, we will automatically analyze the history of electronic messages, such as those communicated via WhatsApp, Messenger and e-mails, of clinically normal participants and AD patients. The literature about the automatic detection of AD using linguistic input has shown that productions from AD patients are automatically distinguishable from productions of normal older adults. Furthermore, case studies about authors who developed AD themselves suggest that their writing style progressively changes as a result of the disease. With respect to existing corpora containing linguistic materials from AD patients, the data that we collect will form a unique corpus; we are not aware of other resources featuring such longitudinal data. In this article, we argue how our project will contribute to the research on AD and discuss our considerations on collecting, processing and sharing the project's data. We also speculate how the data could be used to develop an automated tool for preclinical detection of AD.

Keywords: Alzheimer's Disease, Longitudinal Data, Electronic Messages

1. Introduction

The Mind-It project is an interdisciplinary project comprising collaborative research groups in neuroscience, computational linguistics, and discourse analysis. The project's aim is to use NLP-techniques and linguistic modelling for preclinical detection of Alzheimer's disease (AD), by analyzing the evolution of electronic text messages over time. To develop this technology, a key step in the project is the collection of corpora of electronic messages of AD patients and clinically normal older adults.

The project began in September 2019 and currently we are in the recruitment phase, collecting the electronic messages of French-speaking volunteers. In this article, we first explain the goals of our project in respect to medical AD research. Second, we review literature from the field of computational linguistics on the automatic detection of AD. Third, we present our method for the recruitment of respondents, the construction of the resource, data protection and processing and an example from the corpus. Finally, we present the methods we will use to process the resource for the future development of our early AD-screening tool.

1.1 The Importance of the Preclinical Detection of AD

Alzheimer's disease (AD) is a condition in which the patient's cognitive abilities decline progressively over many years before reaching the dementia stage, at which point the patient loses his or her autonomy in daily life activities. Currently, there is no marketed cure for this disease and many scientists are now turning towards testing preventive strategies to modify the course of the disease (McDade and Bateman, 2017). Upon autopsy, the brains of AD patients are affected by amyloid-β (Aβ) plaques and tau tangles (Nelson et al., 2012). The recent development of in vivo Aβ and tau imaging confirms the hypothesis that Aβ facilitates the development of tau pathology in the neocortex, which in turn leads to cognitive decline (Wang et al., 2016; Hanseeuw et al., 2019).

Growing evidence suggests that Aβ pathology appears 15 to 20 years before the onset of AD dementia (McDade and Bateman, 2017) and that treating amyloid plaques after the onset of dementia does not provide clinical benefits to patients (Selkoe, 2019). Therefore, it would appear that an effective treatment would imply curbing Aβ pathology as soon as possible, before the onset of memory impairment symptoms (McDade and Bateman, 2017).

However, detecting Aβ and tau pathology is expensive and/or invasive. At present moment, there are two reliable methodologies: PET (positron emission tomography) imaging and cerebrospinal fluid (CSF) analysis obtained after lumbar puncture. Both methods have significant drawbacks. PET imaging is very expensive and time consuming. The exam takes half a day for a patient to complete, and requires the injection of radioactive fluids into the blood. CSFs can be painful, are contra-indicated for some patients and include a risk of hospitalization. Above the age of 70, about 20% of the clinically normal population is positive for Aβ pathology and is thus at risk for AD. However, exposing this population to invasive and expensive testing is — especially in the absence of a cure — not advisable.

In conclusion, identifying non-demented older adults with Aβ pathology is crucial for conducting preventive clinical trials, and the development of inexpensive and non-invasive screening tools applicable to the general older population is an important research priority.

1.2 Aims of the Mind-It Project

The aim of our project is to develop a screening tool that detects linguistic decline through a person's history of

electronic conversations. We are developing a computational model based on electronic messages written by AD patients and clinically normal older participants. For every time step in the message history, linguistic performance is automatically evaluated and, in that way, a linguistic performance curve can be established for AD patients and control participants. We expect the AD patients' curve to have a declining slope and hope to be able to match the slope with AD early detection.

Electronic conversation histories are a valuable data source. Contrary to clinical data, that are typically collected once AD is suspected but not before, histories are kept automatically and make it possible to assess the linguistic level of a person before the onset of cognitive problems, provided the history is long enough. This feature allows to estimate whether somebody's linguistic performances are regressing, or whether they are stable, even if the writing does not follow standard conventions. The history of electronic messages allows us to study the influence of AD on linguistic performance at various moments in time, without the necessity for participants to come back to provide us with new data.

2. Literature Review: Automatic AD Detection using Linguistic Data

In this section, we review the literature concerning the automatic detection of AD that relies on the use of written textual data. More precisely, we focus on two types of studies that are important for our project. (1) Studies based on the Pitt Corpus, an important resource shared freely for research purposes. It has a substantial number of participants, with and without AD. Other corpora containing linguistic materials of AD patients exist, but they were often gathered for individual non-reproducible studies and are not shared with the scientific community. (2) NLP studies that rely on longitudinal textual data, from literature writers with and without AD, are also very relevant to our project.

2.1 The Pitt Corpus

A resource that has been very frequently used by computational linguists is the Pitt Corpus, a corpus from the DementiaBank[1] (Becker et al., 1994). The corpus is composed of transcripts and audio files that were gathered for the Alzheimer and Related Dementias Study at the University of Pittsburgh School of Medicine: a longitudinal study that lasted for 5 years from 1983 until 1988 (Bourgeois, 2019). The participants were elderly controls (n = 101), people with probable and possible AD (n = 181), and people with other types of dementia (Becker et al., 1994; Bourgeois, 2019). Language evaluations were part of a series of tests to assess functioning in different cognitive domains: memory, language, visual perception, visual construction, attention, executive functions and orientation.

When the study started, there were 102 subjects enrolled as controls and 204 as AD patients[2]. Subjects with dementia participated in multiple linguistic studies: a fluency task, for which they had to name a maximum number of words on a given theme in one minute (for example, name a maximal number of animals); a recall experiment in which they had to recall a story the experimenter had told them a couple of minutes before; an experiment in which they had to make sentences with one, two or three words given by the investigator; and, finally, the cookie theft picture task (from the *Boston diagnostic examination for aphasia* (Goodglass and Kaplan, 1983)) in which the participants described what was going on in a picture. The control group, for their part, only provided substantial data for the cookie theft picture task. Therefore, it is the cookie theft picture description task that is used most widely in studies that try to automatically detect AD-disease based on linguistic features.

The Pittsburgh cookie theft picture descriptions are used in a large number of studies to build classification systems of AD versus non-AD. The highest accuracy — 0.9742 — using the Pittsburgh cookie theft picture descriptions was obtained by Chen et al. (2019) by using an attention-based hybrid neural network. This is remarkable, especially given the fact that autopsy to confirm AD was only performed on a subset of the AD-participants in the Pitt Corpus. These autopsies showed that a number of participants was falsely diagnosed with AD. Therefore, it is very likely that there is a substantial number of false positives among the 181 AD-tagged participants of the Pitt Corpus.

An interesting study that worked with this same data set is Fraser et al. (2016). They investigated 370 linguistic features, found that around 50 features lead to an optimal model, and made an interpretation of these features, using an exploratory factor analysis. Even though, compared to today's state of the art precision, the accuracy of 81% obtained by Fraser et al. (2016) is not high, nevertheless the feature analysis gives interesting insights into the characteristics of language of AD patients. They found four major factors that play a role in the automatic identification of AD speech: semantic impairment, acoustic abnormality, syntactic impairment and information impairment. We can also cite Karlekar et al. (2018), who obtained an accuracy of 91.1% with a neural network architecture, and Orimaye et al. (2017), who obtained an AUC-score of 0.93 (but not report accuracy).

2.2 Case Studies on AD using Longitudinal Linguistic Data

Several studies have been published in which novels by fiction writers, who were known to (probably) have developed AD, were compared to writers who were considered as a control group. For example, Van Velzen et al. (2014) studied the Type Token Ratio (TTR) and the number of noun and pronoun uses of authors Iris Murdoch, Gerard Reve, Hugo Claus, Agatha Christie, P.D. James and Harry Mulisch. Murdoch was post-mortem confirmed with AD, whereas Reve and Claus

[1]The DementiaBank is part of the larger TALKBank project (MacWhinney et al., 2011). It contains corpora in English, German, Spanish, Mandarin and Taiwanese. It consists of data of AD patients and clinically normal older adults. DementiaBank uses the CHAT format and enables the distribution of audio, video and transcript files.

[2]As clinical AD diagnoses in the 1980s were probable at best, we have to bear in mind that from the whole dataset of participants, 10-20% had other neuropathologies, rather than AD, as the cause of their dementia syndrome.

received a probable AD diagnosis. Agatha Christie was suspected by some scholars to have suffered from AD, but no medical diagnosis was pronounced. Van Velzen et al. (2014) underline the need to consider other models than linear ones, and to test higher order models as well. However, due to the small sample of writers and the absence of a confirmed AD diagnosis — except for Murdoch — the results on the TTR and the noun/pronoun ratio are not very conclusive in distinguishing AD suffering writers from non-AD suffering writers. However, their approach is meaningful for us as they compare text productions from different authors and they therefore depend on inter-individual variation that will have to be taken into account, as it should not be mixed with the AD/non-AD difference.

A second work which is interesting to us is that of Marckx et al. (2018), who performed a study that compared an author with probable AD (Hugo Claus) with an author without AD (Willem Elsschot) on the feature of propositional idea density. For Claus, they included 15 novels and for Elsschot, 11. For each novel, propositional idea density was measured. Propositional ideas can be defined in three ways: 1) predicates, 2) quantifiers and negations, and 3) discourse relations between two propositional ideas. The total number of propositional ideas is the sum of the uses of each of these three factors. Propositional idea density is expressed as the number of propositional ideas per 10 tokens. The measure shows an increase with age for Elsschot and a slight decline with age for Claus. Further analysis should determine whether this metric can be applied to larger samples and also to non-literary genres of corpus, like ours.

2.3 Discussion of Previous Studies

The studies on the Pitt Corpus show that linguistic productions of AD patients are distinguishable from clinically normal older adults. Machine learning techniques, which were employed for these studies, are of interest to the development of screening tools. However, we should note that even though DementiaBank was a longitudinal project that tested the participants every year, this feature is mostly ignored by studies using the Pitt Corpus. For example, two cookie theft picture descriptions from the same participant from two different years, are treated as two descriptions of different participants[3]. Moreover, it should be remembered that the cookie theft picture descriptions are quite a singular corpus and the productions of the participants are very much shaped by the task. Corpora with spontaneous speech, like that of our study, may reveal other aspects about AD. For example, as our corpus contains written e-mails, we could discover more about the influence of AD on discourse structure and coherence.

Antonsson et al. (2019) confirmed that the type of corpus matters. They made an interesting comparison between the cookie theft picture description task and a more complex discourse task. In this second task, participants were asked to describe how they would plan and execute a trip to Stockholm (the participants were all Swedish). The results showed that this task, unlike the cookie theft picture description task, allowed the researchers to discriminate between a group of patients with mild cognitive impairment (n=23) and a group of clinically normal volunteers (n=34).

The literature about authors who developed AD is of significant interest because it provides longitudinal changes in linguistic practices during the preclinical stage of AD (before the onset of overt cognitive symptoms), even though contrary to our corpus, literary work is heavily edited, leaving less traces of AD. However, because of the low number of authors in each study, and often the absence of confirmed AD diagnoses (by autopsy, CSF or PET), the results remain rather anecdotal. For example, it is not clear whether the propositional idea density of Claus diminished because of AD or just because it was the natural evolution of his writing style. It would be interesting to test whether the concept of propositional idea density is meaningful for our corpus as well as more coarse metrics such as the TTR. It is also necessary to evaluate the influence of different features from various linguistic levels (syntax, lexicon, morphology, semantics and discourse) in the same model, without combing them all into one metric.

3. The Mind-It Corpus

In order to build up our corpus, various ethical, methodological and analytical phases are needed. The first phase was the approval of our research protocol by the ethical committee of our research institution and hospital.
The second phase — the current stage of the project — is the collection of data from 30 AD patients and 30 clinically normal older adults. In this section, we will first go through the considerations of the ethical committee, our participants, and how participants give their informed consent. Then, we describe the current phase in more detail: how we recruit participants and how we protect and process their data. At the end of this section, we give an example of messages from our corpus to illustrate how AD shows in longitudinal data of one patient. In the following section, we explain how this first version of the corpus can be used for the development of an early AD detection tool and how we will eventually assess the performance of this tool.
In Figure 1, all the phases of the Mind-It project are represented in a diagram.

3.1 Considerations of the Ethical Committee

The protocol of the Mind-It project was approved on the 17th of September 2019 by the ethical committee of Université catholique de Louvain (UCL) and the academic hospital Cliniques universitaires Saint-Luc in Brussels, under the registration number B403201941006.

One important condition for the approval of the protocol was to block the access to patients' medical data from the linguistic team in charge of the project and to disable access to non-anonymous content of electronic messages to the medical team in charge. So, the healthcare

[3]We should nevertheless remark that not every participant has multiple interventions in the corpus. Indeed, from one year to another the dropout of participants was quite high.

professionals cannot read their patients' messages and linguists do not have access to the medical records of the patients.

A second important point is that our corpus is made up of electronic dialogues between the participants and all of the addressees. Consequently, only messages sent by the participant are kept, and received messages from their correspondents are deleted from the corpus.
From a discursive point of view, it would be interesting to work on the conversation as a whole, as AD features may emerge from the textual context — and even co-text — but participants do not have the right to transfer the copyright of messages written by a third party.

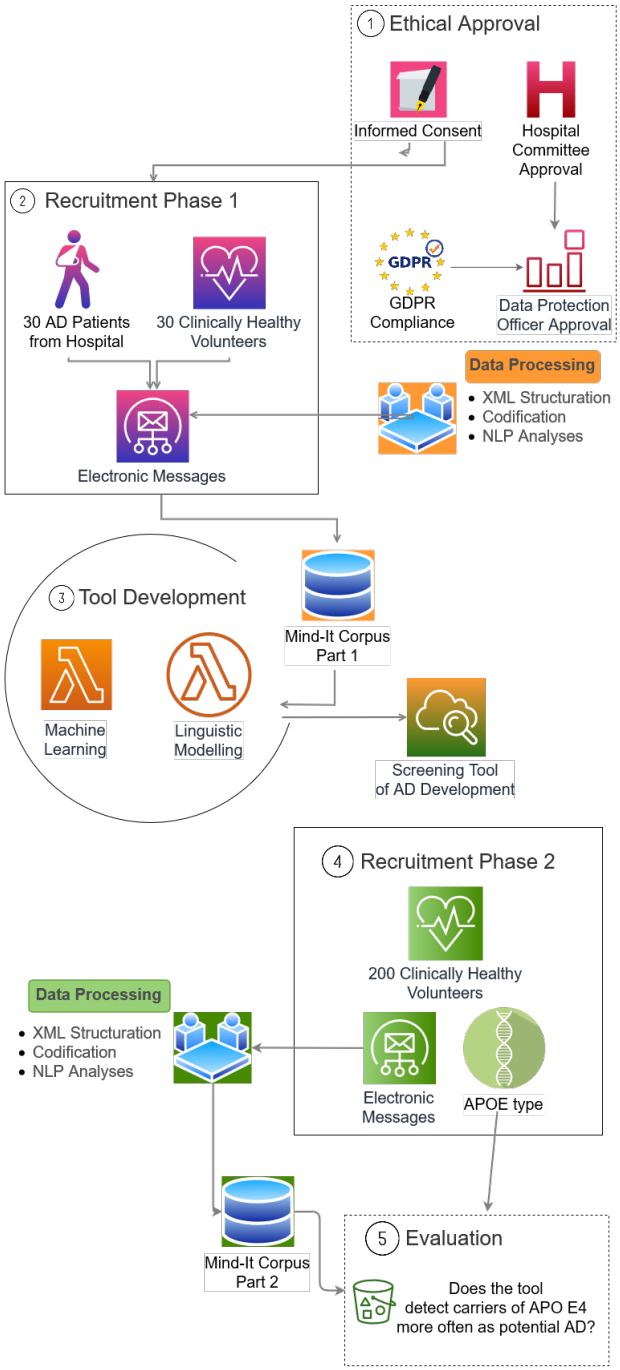

Figure 1: The phases of the Mind-It project.

3.2 Informed Consent

Participants are invited to read and sign the Mind-It project's informed consent form before the start of the collection. On this form, they transfer copyright on their data to the UCL. The informed consent states that the data cannot be used commercially and is only for research purposes at our institution, and that participants' privacy will be guaranteed. Furthermore, it explains to participants that they have a withdrawal right that enables them to withdraw at any point from the project without any explanation. If a patient is under guardianship and wishes to participate, the legal guardian needs to sign the informed consent.

3.3 Participants

Since September 2019, we have been collecting data from patients with prodromal AD and mild AD dementia as well as from clinically normal older adults. In the first phase of the project, our objective is to recruit 30 participants for each category.

AD patients are recruited from the academic hospital Cliniques universitaires Saint-Luc. They have been formally diagnosed with AD either by means of a cerebrospinal fluid punction in which Aβ and tau were searched for or PET imaging. They are followed by the hospital's memory clinic and have undergone a neuropsychological assessment to monitor their cognitive abilities. Furthermore, for these patients, their Apolipoprotein E (APOE) genotype was established. Some expressions of this gene have been related to an enhanced risk of developing AD (Hauser and O Ryan, 2013). However, it is impossible to say whether somebody will develop AD based on their APOE genotype only: people having expressions for a higher risk don't necessarily develop AD and people with a low risk expression can still develop it.

Older volunteers are recruited through two channels: either we ask spouses that often accompany AD patients to the clinic, or we recruit via the University for the Elderly linked to the UCL. In contrast to the AD patients, we do not dispose of the neuropsychological evaluations of these volunteers. Therefore, we ask them (1) to certify they do not have major cognitive impairment and (2) whether we can evaluate their APOE genotype by the means of a simple blood test. The results of APOE testing is not provided to the volunteers as it is only a risk evaluation, and no reliable conclusions can be drawn as to whether a specific individual will develop AD or not.

At a later stage of the project — the evaluation of our early detection tool that we aim to develop — we plan to recruit a maximum number of elderly people without an AD diagnosis. We will elaborate on this in section 4.3.

3.4 Data Collection

After informed consent is given, we ask the participants to fill in a socio-demographic form which includes questions about their age, education, level of activity and other health conditions that may have an impact on the language and or writing (sight, arthrosis, etc.). This information may have relevance for the evaluation of the data.

The collection of the participants' electronic messages constitutes the most important part of the research project. We are interested in various types of electronic messages, coming from different applications and devices (mainly smartphones and computers). As far as applications are concerned, we gather data from any electronic message service, including Gmail, Outlook, Messenger, WhatsApp, Skype, Viber, and Telegram. Most of these services offer export tools that enable us — through varying levels of ease — to collect all messages that have been sent, to get a maximal history. For each application, a specific and distinct protocol has been drafted by our team, following each application's technical specificity.

The data collection may happen in the presence of the participants and the collector responsible, or by the participants themselves, based on the type of electronic messages they want to donate and their confidence in their ability to copy the messages correctly and transfer them to us. If the participant needs assistance, the person in charge meets them at the hospital, the university or the participant's residence. We encourage participants to donate their entire history of sent messages and not making a selection themselves of what to donate and what not, but participants are free to remove conversations or messages, if they do not feel like sharing them. So far, the large majority of participants shared all their messages.

3.5 Data Protection

Our data collection ensures GDPR (General Data Protection Regulation) compliance, which is needed for research projects collecting human data. This has received the agreement of the official Data Protection Officer from the UCL. Data is stored on protected servers of the university. The data will be semi-automatically codified before its processing by the linguistic team: sensitive information such as names, surnames, (e-mail) addresses, phone numbers, and bank account numbers will be removed.

Example (1) from the *Vos Pouces pour la Science* corpus — a corpus of electronic conversations in French — (Panckhurst and Cougnon, 2019) illustrates the type of codification we plan to apply to our data.

(1) {name}, le numero d'{name} qui est a {address} et espere te voir, {number} Bisous!!! PS: j'ai pas ton numero francais!!
{name}, the number of {name} who is at {address} and hopes to see you, {number} Kisses!!! PS: I do not have your French number!!

3.6 Data Processing

The first step of data processing is to parse the electronic messages from different messaging platforms and to save them in an exploitable homogeneous format. For each participant we will create an XML-file, in which every message is a node, associated with some meta data such as the timestamp and the platform (e-mail, WhatsApp, etc.) source. In this XML-file, we will also include the information from the socio-demographic questionnaire, but no medical data other than whether the participant is AD or clinically normal.

Medical data will be stored in protected electronic medical records. After pseudo-anonymization, medical information such as clinical diagnoses and APOE genotyping will be extracted into protected research files. The inclusion of linguistic parameters obtained from the XML file to this pseudo-anonymized research file will only be made by authorized personnel from the university hospital. Researchers from both the hospital and the university will only be granted access to this pseudo-anonymized data file that will not include access to raw messages.

3.7 Data Sharing

Because of our participants' privacy, we cannot freely share all the collected data outside of the university. The corpora, especially e-mail corpora over several years, are of such a considerable size that manual codification is not a viable solution. As participants' privacy must be guaranteed, we cannot use a (semi)-automatic codification that may leave some private information in the corpus. However, as we are convinced of the necessity of open-source and replicable research results, we will distribute all collection details (consent, form, ethical and GDPR material) as well as the (automatic) linguistic analyses we will run to process the data, such as part of speech tagging and syntactic parsing. Currently, we are also investigating whether it is possible to release some subparts of the corpus after manual correction of the automatic codification.

3.8 Example

In this subsection, we present two extracts from our corpus from the emails of a patient diagnosed with AD. We want to illustrate the idea that the progression of AD can be visible when we look at longitudinal data, such as an e-mail corpus. In the examples, bold font is used to mark parts of the message that do not follow French writing conventions and between brackets we give the correct form.

(2) Message sent in July 2013:

Bonjour {name},
Hello {name},

Je n'ai finalement pas pu vous attendre hier soir car votre réunion a été importante et longue**!** ***[exclamation mark should be preceded by a white space]***
In the end, I could not wait for you yesterday evening because your meeting was important and long!

Pour votre information, en partant hier soir {name} m'a dit que demain à la **pause café** ***[pause-café]*** vers 10h, il y aura une petite fête d'adieu pour {name} et {name}.
For your information, when I left yesterday evening {name} said to me that tomorrow during the coffee break around 10a.m. there will be a little farewell party for {name} and {name}.

A demain,
See you tomorrow,

{name}
{name}

This e-mail does practically not contain mistakes regarding the writing conventions. However, in example (3) that was written three years later by the same patient, we see mistakes in punctuation, spelling and the use of colloquial language, whether the tone of the message is rather formal.

(3) Message sent in October 2016 :

Comment **allez vous [allez-vous] ?? [One question mark too much] La santé est bonne ? [colloquial language]**
C'est vraiment dommage que vous ne soyez **plu [plus]** là.
How are you?? Your health good ? It is really a shame that you are not there anymore.

J'ai une **question,certainement [question, certainement]** vous pouvez m'aider à résoudre.
I have a question you can certainly help me to answer.

Concerne [Concernant] votre lettre du {date} relative à la facture intermédiaire pour les travaux de renouvellement de l'ascenseur.
About your letter of the {date} concerning the intermediary bill for the renovation works of the elevator.

Vous réclamiez deux versements :
le premier de 1.285,52 **€([missing white space]** et pour cela je trouve le débit sur mon extrait de compte le {date}) ; mais dans la lettre vous indiquez de verser pour la fin de la semaine suivante 514,21 **€, [.]**
You claimed two payments: the first of 1,285.52 € (and for that one I find the debit transaction in my account statement on the {date} ; but in the letter you wrote that you would transfer 514.21 € by the end of next week,

Pour ce versement je ne trouve rien. **Cela vous rappelle quelque chose ? [colloquial language]**
I do not find a trace of this payment. Does it remind you of something?

Je vais aussi **à [le]** demander à ma banque, mais en principe j'ai encore tous **le [les]** extraits.
I will also ask my bank, but normally I still have all the extracts.

Merci pour tout le travail que vous avez fait (et c'est un grand dommage que vous ne soyez plus là) **[missing period]**
Thank you for all the work you did (and it is really a shame that you are not there anymore).

Bonjour à Madame. (et à une prochaine fois).
Give my regards to Mrs (and see you next time).

{name}
{name}

These two extracts show that our corpus contains data that make it possible to assess the linguistic level of a participant over time. Compared to corpora gathered in a clinical setting, this corpus contains linguistic output of a participant before and after their AD diagnosis. By comparing participants to anterior versions of themselves, it can be estimated whether a lower linguistic level can be attributed to AD or not.

4. A Tool for Early AD Detection

As our corpus is still in the collection phase, we have not yet started on the development of the tool for the early detection of AD. Nevertheless, we are already able to discuss the considerations we have about it thus far.

4.1 NLP Analyses

In order to use our corpus for the development of our tool, we want to apply different types of automatic linguistic analysis to it. We plan to perform syntactic analysis, such as part-of-speech tagging and constituency — or dependency — parsing (Ribeyre et al., 2016; Coavoux and Crabbé, 2017). We also want to consider automatic semantic analyses. For example, Ribeyre et al. (2016)'s parser provides surface syntactic analysis, as well as a 'deep' syntactic analysis: not only are surface grammatical functions annotated, but also the semantic predicate argument structures. We are also interested in analyses of discourse structure (Braud and Denis, 2013) to see whether discourse coherence is affected by AD.
An important challenge will be to adapt existing systems to our genre of data. As many available tools were developed on manually annotated corpora consisting of journalistic texts, the question arises whether their performance on different types of electronic messages from our corpus will be of sufficient quality. Furthermore, it should be kept in mind that our corpus is in French and here that there are fewer resources available than for English (even if, amongst all languages of the world, French is quite well represented in NLP).

4.2 Type of Model

The type of statistical model we want to use for the tool is heavily dependent on different criteria of the project: performance on the early detection of AD, the interpretability of the model and the guarantee of privacy of the electronic messages. When we consider the first aspect, looking at studies performed on the Pitt Corpus, it appears that a neural network architecture will lead to the highest performance in terms of AD detection. But, if we consider the two other aspects, we are not sure that the neural network will be the best choice. As neural networks have an internal feature selection, it can be difficult to understand what, in the electronic messages of AD-patients, distinguishes them from the normal older adults. This is also quite well illustrated by the literature about the cookie theft picture task description: articles, such as the one of Fraser et al. (2016), offer a far better understanding of linguistic markers of AD than articles with a state-of-the-art performance on the data set (Chen et al., 2019). Our third criterion, the guarantee of privacy,

should also be considered. Recent studies have demonstrated that sensitive, private information from the training corpus can be (partially) recovered from the hidden layers of deep neural networks (Coavoux et al., 2018; Carlini et al., 2019). If we decide to develop a tool based on a neural architecture, careful consideration should be given as to how the training can be adapted to avoid the possibility of recovering private information from our model and how the model should be distributed and protected. In particular, we have to evaluate whether the automatic codification of the training corpus is sufficient.

Because of the criteria of interpretability and privacy, we are also considering developing other types of computational models, for example (generalized) mixed effects models (Agresti, 2002). The advantage is that these models have a high interpretability: they can estimate the effect size of specific linguistic features of AD. Moreover, as the feature selection is manual for this type of model, there is no risk of privacy issues.

4.3 Evaluation

There are two ways in which we want to evaluate our tool. The first is a rather classical method: cross validation, to evaluate the accuracy and robustness of the model. The second method is less conventional: we want to recruit more participants (our objective is 200), older than 60, who have not been diagnosed with AD. It is crucial that these participants not only give their electronic message histories, but also participate in the blood test of Apolipoprotein E (APOE). We want to run our tool on their messages and see whether there is a statistical relation between having an increased genetic risk of developing AD and the outcome of our screening tool. If there is, it will be an important argument that our tool could help to detect AD in the preclinical stage. We plan to organize a different collection campaign with motivational prizes to achieve this aim.

4.4 Ethical Aspects

If our screening tool would be successful, special consideration should be given to the ethical aspects of its use. We aim for a tool that can only be used after one gives their consent and delivers their own electronic message history. We have absolutely no intention of developing a tool that runs in the background of devices or other applications and that keeps statistics over one's linguistic performances and estimates continuously their risk for AD. Our purpose is to make this tool available in a clinical framework: if the tool suspects AD, it is crucial to propose medical examination. The tool can absolutely not replace the medical exams that are used to diagnose AD; it has merely the purpose of a screening a device. Moreover, the electronic message history of people using the tool should not be stored, except if the participant explicitly agrees to use their data to enhance future performances. In that case, the data should by no means be shared with third parties.

5. Conclusion

As far as we know, there hasn't yet been a project aimed at developing a longitudinal model of the progression of AD evidenced in written text, other than the studies of authors that are presumed to have suffered from AD. However, as only productive writers build up a rich body of literary work over their life time, these models are not applicable to a wider public. We propose to use smartphone data (chat conversations) and emails as a source of longitudinal data. As more and more people have smartphones, it is likely that our model can apply to a large population. If the longitudinal model is able to screen for patients in a preclinical stage of the disease, it could contribute significantly to the early detection of the disease and therefore to the recruitment of participants in drug studies that only focus on patients who do not yet present cognitive impairment.

6. Acknowledgements

We thank the anonymous reviewers for their comments. The Mind-It project is financed by the SAO-FRA grant from the *Stichting Onderzoek Alzheimer / Fondation de Recherche Alzheimer*.

7. Bibliographical References

Agresti, A. (2002). *Categorical Data Analysis*, volume 482. John Wiley & Sons.

Antonsson, M., Fors, K. L., and Kokkinakis, D. (2019). Discourse in mild cognitive impairment. *ExLing 2019*, 25:21.

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.

Bourgeois, M. (2019). Dementiabank progress and future plans. U R L : http://gandalf.talkbank.org/ symposium/0pptx/Bourgeois.pptx, 6.

Braud, C. and Denis, P. (2013). Identification automatique des relations discursives «implicites» à partir de données annotées et de corpus bruts. *TALN-RÉCITAL 2013*, pages 104–117.

Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., and Song, D. (2019). The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pages 267–284.

Chen, J., Zhu, J., and Ye, J. (2019). An attention-based hybrid network for automatic detection of Alzheimer's disease from narrative speech. *Proc. Interspeech 2019*, pages 4085–4089.

Coavoux, M. and Crabbé, B. (2017). Représentation et analyse automatique des discontinuités syntaxiques dans les corpus arborés en constituants du français. In *24e Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, pages 77–92.

Coavoux, M., Narayan, S., and Cohen, S. B. (2018). Privacy-preserving neural representations of text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1–10.

Fraser, K. C., Meltzer, J. A., and Rudzicz, F. (2016). Linguistic features identify Alzheimer's disease in

narrative speech. *Journal of Alzheimer's Disease*, 49(2):407–422.

Goodglass, H. and Kaplan, E. (1983). Boston diagnostic examination for aphasia. *Philadelphia: Lea and Febiger*.

Hanseeuw, B. J., Betensky, R. A., Jacobs, H. I., Schultz, A. P., Sepulcre, J., Becker, J. A., Cosio, D. M. O., Farrell, M., Quiroz, Y. T., Mormino, E. C., et al. (2019). Association of amyloid and tau with cognition in preclinical Alzheimer disease: A longitudinal study. *JAMA neurology*.

Hauser, P. and O Ryan, R. (2013). Impact of apolipoprotein e on Alzheimer's disease. *Current Alzheimer Research*, 10(8):809–817.

Karlekar, S., Niu, T., and Bansal, M. (2018). Detecting linguistic characteristics of Alzheimer's dementia by interpreting neural models. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 701–707, New Orleans, Louisiana, June. Association for Computational Linguistics.

MacWhinney, B., Fromm, D., Forbes, M., and Holland, A. (2011). Aphasiabank: Methods for studying discourse. *Aphasiology*, 25(11):1286–1307.

Marckx, S., Verhoeven, B., and Daelemans, W. (2018). The Claus case: Exploring the use of propositional idea density for Alzheimer detection. *Computational Linguistics in the Netherlands Journal*, 8:66–82.

McDade, E. and Bateman, R. J. (2017). Stop Alzheimer's before it starts. *Nature News*, 547(7662):153.

Nelson, P. T., Alafuzoff, I., Bigio, E. H., Bouras, C., Braak, H., Cairns, N. J., Castellani, R. J., Crain, B. J., Davies, P., Tredici, K. D., et al. (2012). Correlation of Alzheimer disease neuropathologic changes with cognitive status: a review of the literature. *Journal of Neuropathology & Experimental Neurology*, 71(5):362–381.

Orimaye, S. O., Wong, J. S., Golden, K. J., Wong, C. P., and Soyiri, I. N. (2017). Predicting probable Alzheimer's disease using linguistic deficits and biomarkers. *BMC bioinformatics*, 18(1):34.

Panckhurst, R. and Cougnon, L-A. (2019). Youth Digital Practices: results from Belgian and French projects. In *TechTrends*, pages 1-10.

Ribeyre, C., de la Clergerie, E. V., and Seddah, D. (2016). Accurate deep syntactic parsing of graphs: The case of french. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3563–3568.

Selkoe, D. J. (2019). Alzheimer disease and aducanumab: Adjusting our approach. *Nature reviews. Neurology*.

Van Velzen, M. H., Nanetti, L., and De Deyn, P. P. (2014). Data modelling in corpus linguistics: How low may we go? *Cortex*, 55:192–201.

Wang, L., Benzinger, T. L., Su, Y., Christensen, J., Friedrichsen, K., Aldea, P., McConathy, J., Cairns, N. J., Fagan, A. M., Morris, J. C., et al. (2016). Evaluation of tau imaging in staging Alzheimer disease and revealing interactions between β-amyloid and tauopathy. *JAMA Neurology*, 73(9):1070–1077.