Received: 16 June 2020 | Revised: 13 December 2020 | Accepted: 6 January 2021 | Published online: 16 February 2021

DOI: 10.1002/trc2.12147

Alzheimer's & Dementia
Translational Research & Clinical Interventions

SHORT REPORT

Multilingual automation of transcript preprocessing in Alzheimer's disease detection

Frédéric Abiven | **Sylvie Ratté**

Department of Software and IT Engineering, École de technologie supérieure, Montreal, Quebec, Canada

Correspondence
Frédéric Abiven, Department of Software and IT Engineering, École de technologie supérieure, 1100 Notre-Dame St W, Montreal, QC H3C 1K3, Canada.
E-mail: frederic.abiven.1@ens.etsmtl.ca

Abstract

Introduction: Analyzing linguistic functions can improve early detection of Alzheimer's disease (AD). To date, no studies have focused on creating a universal pipeline for clinical transcript preprocessing.

Methods: This article presents a simple and efficient method for processing linguistic and phonetic data, sequencing subproblems of cleaning, normalization, and measure extraction tasks. Because some of these tasks are language- and context- dependent, they were designed to be easily configurable, thus increasing their scalability when dealing with new corpora.

Results: Results show improved performances over previous studies in this time-consuming preprocessing task. Moreover, our findings showed that some discursive markers extracted from transcripts revealed a significant correlation (>0.5) with cognitive impairment severity.

Discussion: This article contributes to the literature on AD by presenting an efficient pipeline that allows speeding up the transcripts preprocessing task. We further invite other researchers to contribute to this work to help improve the quality of this pipeline (https://github.com/LiNCS-lab/usAge).

KEYWORDS
Alzheimer's disease, discursive markers, early detection, linguistic features, phonetic features, pipeline, transcript preprocessing

1 | INTRODUCTION

Monitoring and detecting Alzheimer's disease (AD) at an early stage is becoming more crucial as the number of people affected by the disease increases rapidly every year. Currently, nearly 50 million people are living with AD globally, and that number is expected to reach 152 million by. Many studies have covered computer-based approaches to evaluating and monitoring cognitive functions and detecting AD at an early stage.[2,3] The Cookie Theft picture description task is widely used to monitor and detect the disease. In this study, we analyze transcripts and audio clips from the Pitt Corpus[4] (English) and the CRIUGM (Centre de recherche de l'Institut universitaire de gériatrie de Montréal) Corpus (Quebec French),[5] as listed in Table 1. Extracting valuable measures can be difficult when working with multilingual datasets. Therefore, this work presents a multilingual pipeline approach that preprocesses and extracts multiple linguistic and phonetic characteristics from data. To evaluate our approach, we compared the results of both datasets.

TABLE 1 Distribution of interviews used for experimentation

Corpus name	Language	Criteria	Diagnosis/type	Samples
CRIUGM	French	<40 y/o	Healthy young	26
		>50 y/o	Old	29
Pitt Corpus	English	MSSE	HC	242
			AD or MCI	300

Abbreviations: AD, Alzheimer's disease; HC, healthy control; MCI, mild cognitive impairment; MMSE, Mini-Mental State Examination.

RESEARCH IN CONTEXT

1. **Systematic review**: While many studies on computer-based approaches to the early detection of Alzheimer's disease (AD) in a picture description task context have shown great potential over the past few years, most of them cover a specific language. Generally, studies analyzing transcripts based on patients' interviews tend to restrict their research to a specific cognitive task.
2. **Interpretation**: We developed a multilingual and context-independent pipeline for transcript preprocessing (https://github.com/LiNCS-lab/usAge), which extracts a variety of linguistic and phonetic measures that can eventually be used to monitor and detect AD at an early stage.
3. **Future directions**: Transcript preprocessing plays a key role in extracting linguistic measures, as it holds valuable information for cognitive assessment. Further research could focus on understanding how linguistic functions are altered differently in patients with different spoken languages.

2 | METHODS

In this work, we present a methodology based on a pipeline architecture for processing transcripts. This allows the division of the work into subprocesses, which makes it easier to approach the multilingualism factor. Each subprocess is seen as a single entity that can be adapted to different languages and contexts. The pipeline is divided into the following six main modules: typographic normalization and cleaning, part-of-speech (POS) tagging, POS adjustment, POS distribution measurement, linguistic measurement, and phonetic measurement. Multilingual modules are identified in blue, as illustrated in Figure 1. We will go through each of the modules to explain how they contribute to transcript preprocessing.

2.1 | Typographic normalization and cleaning

Working with transcripts carries multiple challenges due to the sparsity of related norms. Transcripts may appear in different formats, such as plain text files or transcription files (.cha). Also, different discursive marker norms can be used in annotating transcripts as most are produced by hand. Typographic errors could also be injected into transcripts as they are human-made. To tackle this problem, the cleaning and normalization task is easily adjustable with configuration files, using a rule-based approach. This allows adaptation of the process to match different languages and different interview contexts, as we can specify new rules. The process thus cleans transcripts and extracts discursive markers, which have been shown to correlate highly with the

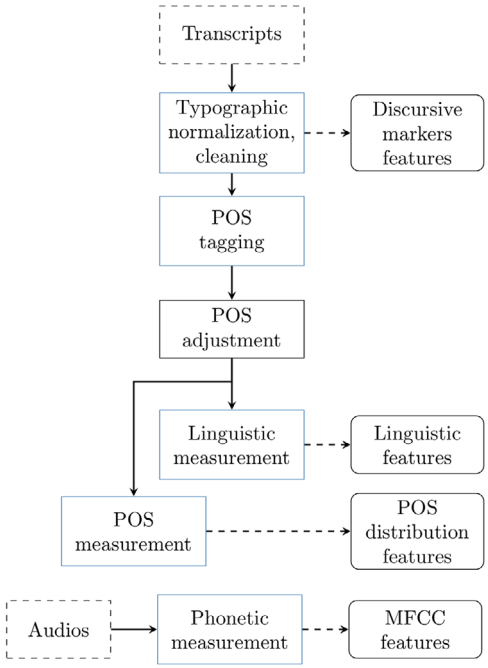

FIGURE 1 Transcript preprocessing pipeline architecture. MFCC, mel-frequency cepstral coefficients; POS, parts of speech

disease. In fact, they are widely used in the best performing predictive models to detect AD in English and in French, as shown in Table 1 (respectively 6/10 and 10/10).

2.2 | POS tagging

POS tagging tools have proven their effectiveness in recent years and are now widely used in natural language processing tasks. In our work, we used FreeLing 4.0 to analyze and tag transcripts, because it supports many different languages,[6] although its flexibility in tagging words and tagging norms may vary from one language to another. Authors of FreeLing have reported >95% accuracy on journalistic texts; this limitation regarding the training of POS-taggers is addressed in Section 4. As an addition to this module, we therefore converted tags to a universal form, allowing the following modules to analyze and manipulate transcripts from various corpora. This task was tested on both English and French transcripts but may be used for numerous other languages.[6]

2.3 | POS adjustment

Because POS tags are statistically determined, some annotation errors may be introduced into transcripts. This module consists mainly in finding the most common mistakes and updating them to the correct form programmatically. It evaluates and analyzes the tags, thus allowing improvements in the quality of the results in the following modules. However, this module must be adapted only once for each language because it depends on a language's structure and rules. In our work, we adapted it for English and French tags.

2.4 | POS distribution measurement

To measure the distribution of POS tags, the frequency and ratio of the following tags were evaluated: adjectives, conjunctions, nouns, prepositions, verbs, and auxiliary verbs. Because the POS tagging module universalizes tags, this process can be applied to different languages.

2.5 | Linguistic measurement

Linguistic characteristics were automatically extracted within this module. We used the most common linguistic measures from previous works, and which have shown a significant correlation with the disease[7] (e.g., Brunet's index, type-token ratio, Honore's statistic). Because these characteristics are based on straightforward distribution of words and lemmas, this module is language-independent.

2.6 | Phonetic measurement

For phonetic characteristics, we used the python_speech_features 0.6 tool to estimate the first 13 mel-frequency cepstral coefficients (MFCCs).[8] We then estimated the mean, kurtosis, skewness, and variance of those values. Audio files of interviews normally consist of a patient and an interviewer speaking, so we segmented the audio to keep only the patient. In future work, a speaker diarization could be done to extract the patient's voice and thus increase the accuracy of the phonetic measurements.

3 | RESULTS

To understand all our linguistic and phonetic measures and how they interact with AD, we performed a correlation analysis. All in all, we extracted 100 features, separated into four different categories: discursive markers, POS distribution, linguistic characteristics, and phonetic characteristics. We also included information coverage measures that were presented in the work of Hernández-Domínguez.[7] We then ran a feature selection process to extract the most valuable features. With the selected features, we trained different predictive models and evaluated their performance with a 10-fold cross-validation, as presented in Table 2. Finally, we analyzed the correlation of the extracted measures with the disease.

3.1 | Discursive markers

Discursive markers have demonstrated their ability to distinguish healthy controls from AD patients quite remarkably. One of the most

TABLE 2 Average AUC on 10-fold cross-validation models with different feature type combinations (baseline = decision tree classifier)

CRIUGM[a] Corpus (French)				
Feature types	**Model**	**AUC**	**F-score**	**F-score baseline**
Cov-ling-phon-	Svm	0.92	0.91	0.83
Cov-phon-pos-	Svm	0.92	0.97	0.91
Cov-ling-phon-pos-	Svm	0.90	0.93	0.81
Markers-ling-phon-pos-	Svm	0.89	0.96	0.77
Cov-phon-	Svm	0.89	0.90	0.91
Markers-ling-	Svm	0.88	0.89	0.85
Markers-cov-ling-	Svm	0.88	0.86	0.82
Markers-cov-ling-	Rfc	0.86	0.86	0.82
Markers-ling-phon-	Rfc	0.86	0.90	0.81
Markers-cov-phon-pos-	Svm	0.86	0.93	0.86
Pitt Corpus (English)				
Feature types	**Model**	**AUC**	**F-score**	**F-score baseline**
Markers-cov-phon-pos-	Svm	0.76	0.79	0.68
Markers-cov-	Svm	0.74	0.77	0.67
Markers-cov-ling-pos-	Svm	0.74	0.77	0.69
Markers-cov-ling-	Svm	0.73	0.77	0.66
Markers-phon-pos-	Svm	0.73	0.76	0.69
Markers-cov-pos-	Svm	0.73	0.76	0.69
Markers-cov-phon-	Svm	0.73	0.75	0.68
Markers-ling-phon-pos-	Svm	0.73	0.75	0.70
Markers-ling-pos-	Svm	0.72	0.74	0.67
Markers-cov-ling-phon-	Svm	0.72	0.75	0.69

Abbreviations: AUC, area under the curve; cov, information coverage features; ling, linguistic features; markers, discursive markers features; phon, phonetic features; POS, parts of speech; pos, POS distribution features.

[a]Centre de recherche de l'Institut universitaire de gériatrie de Montréal

correlated features with these markers is the number of false starts in both English and French corpus (respectively, 0.26 and 0.62). We hypothesize that patients with AD tend to forget how to describe an object or a person, which forces them to retrace their sentences. Also, we found an inverse correlation with the number of synonyms extracted from transcripts in both languages (respectively, −0.13 and −0.31). This could be explained by the fact that AD patients have a smaller vocabulary variety when describing an image. Finally, the number of repetitions detected in both corpora correlates highly with the disease (respectively, 0.35 and 0.37), which is consistent with previous studies.[9,10]

3.2 | POS distribution

For the POS tags distribution, auxiliary verb frequencies were not correlated in the same way in English and in French. We found that in French, the correlation was positive (0.28), while in English it was negative (−0.16). This could be due to the fact that auxiliary verbs cannot necessarily be translated in the same way between the languages (e.g., Je suis allé à l'école: I went to school) and therefore, measures may vary. Similarly, conjunctions and adjectives did not have the same type of correlation between English and French. On the other hand, we found that AD patients tend to use fewer nouns in both languages, which correlates with previous findings.[11] That being said, a POS distribution should be considered and analyzed in each language separately, because it does not necessarily have the same representation in each case.

3.3 | Linguistic characteristics

For the Pitt Corpus, lexical richness correlations were mostly consistent with previous studies.[7] With the CRIUGM dataset, most measures were inconsistent with the results obtained with the Pitt Corpus, and indeed, were sometimes highly correlated with the disease (e.g., Yule's characteristic K [0.44]). We believe that this could be due to the size of the dataset, which is very small, compared to the English dataset. Nonetheless, this module may be considered a benchmark, because the results match those of the same experiment conducted on the Pitt Corpus.[7]

3.4 | Phonetic characteristics

Considering phonetic characteristics, results with the Pitt Corpus are relatively consistent with previous studies.[7] There may have been some differences in correlation values due to the fact that we segmented the audio to remove the interviewer's voice. For the CRIUGM dataset, some of the MFCCs' mean, skewness, and variance values were highly correlated with the disease (>0.4). Again, those high correlations might be explained by the size of the dataset and the manual audio segmentation task, which could bias the results.

3.5 | Modeling

For both corpora, we tested different combinations of feature types, which showed discursive markers to be the most common feature type found in the best predictive models overall. With the Pitt Corpus, our best model had an average area under the curve (AUC) of 76%, which is relatively consistent with previous studies.[7,12] Looking at the CRIUGM Corpus, our best model had an average AUC of 92%. This result, which is significantly high, may be explained by the very small dataset size and the high correlation found in multiple features.

4 | DISCUSSION

This work contributes in many ways to improving the quality and efficiency of transcript and audio preprocessing to extract measures that characterize linguistic and phonetic functions.[a] Furthermore, we expand its use by making the processing adaptable to many different languages. Results demonstrate its consistency with previous studies, as well as with a new cohort of French participants. Although we suspect that FreeLing POS-taggers are not entirely reliable for speech data in various languages, the results were sufficiently reliable to build the pipeline. In a future version, we intend to replace this library with spaCy's library, which has been trained on a wider type of texts (including speech).[13] Further research could focus on including languages with different structures and rules, as that could expand its usage. We would also like to include the information coverage measure extraction as part of a new module in our pipeline, as it has proven its capacity to significantly distinguish AD patients from healthy controls.[7] Finally, we believe it would be interesting to compare results between proportionate datasets of different languages to evaluate how the disease may affect cognitive functions in patients differently.

ACKNOWLEDGMENTS

The research presented in this paper was financially supported by NSERC (Natural Sciences and Engineering Research Council of Canada) RGPIN-2018-05714 and approved by the ethical comity of École de technologie supérieure (H20170506).

CONFLICTS OF INTEREST

The authors have declared that no conflicts of interest exists.

REFERENCES

1. Alzheimer's Disease International. *World Alzheimer Report 2019: Attitudes to dementia*. London, UK: Alzheimer's Disease International; 2019.
2. Weiner J, Frankenberg C, Schroder J, et al. Speech reveals future risk of developing dementia: predictive dementia screening from biographic interviews. In: *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. New York, NY, USA: IEEE; 2019:674-681.
3. You Y et al. Predicting dementia risk using paralinguistic and memory test features with machine learning models. In: *2019 IEEEHealth-*

[a] However, a team from our laboratory is currently dedicated to improving automatic transcription systems in the limited context of image description tasks.

care Innovations and Point of Care Technologies, (HI-POCT). New York, NY, USA: IEEE; 2019:56-59.
4. Becker JT. The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. *Arch Neurol*. 1994;51(6):585-594.
5. Brambati, S *Centre De Recherche De l'Institut Universitaire De Gériatrie De Montréal. Corpus Clinique Francophone D'une Tâche De Description D'Image*. Montréal, QC, Canada; 2017.
6. Padró L and Stanilovsky E. *Freeling 3.0: Towards Wider Multilinguality*. LREC 2012. 2012.
7. Hernández-Domínguez L, Ratte S, Sierra-Martinez G, et al. Computer-based evaluation of Alzheimer's disease and mild cognitive impairment patients during a picture description task. *Alzheimer's Dement*. 2018:260-268.
8. . *Python Speech Feature extraction*, & Lyons, J , . : ; 2017. https://pypi.org/project/python_speech_features/.
9. Guinn CI and Habash A. Language analysis of speakers with dementia of the Alzheimer's type. In: *2012 AAAI Fall Symposium Series*. Vancouver, Canada: AAAI; 2012.
10. Pompili A, Abad A, de Matos DM, et al. Pragmatic aspects of discourse production for the automatic identification of Alzheimer's disease. In: *IEEE Journal of Selected Topics in Signal Processing*. New York, NY, USA: IEEE; 2020;14(2):261-271.
11. Jarrold W, Peintner B, Wilkins D, et al. Aided diagnosis of dementia type through computer-based analysis of spontaneous speech. P Resnik R Resnik & M Mitchell In: *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*; 2014:27-37). Baltimore, Maryland, USA: Association for Computational Linguistics.
12. Fraser KC, Meltzer JA, Rudzicz F. Linguistic features identify Alzheimer's disease in narrative speech. *J Alzheimer's Dis*. 2016;49(2):407-422.
13. Honnibal M and Montani I. SpaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. 2017. https://spacy.io/

How to cite this article: Abiven F, Ratté S. Multilingual automation of transcript preprocessing in Alzheimer's disease detection. *Alzheimer's Dement*. 2021;7:e12147. https://doi.org/10.1002/trc2.12147