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 Corpus4 (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.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

TABLE 1  
Distribution of interviews used for experimentation  

<table>
<thead>
<tr>
<th>Corpus name</th>
<th>Language</th>
<th>Criteria</th>
<th>Diagnosis/type</th>
<th>Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRIUGM</td>
<td>French</td>
<td>&lt;40 y/o</td>
<td>Healthy young</td>
<td>26</td>
</tr>
<tr>
<td>Pitt Corpus</td>
<td>English</td>
<td>&gt;50 y/o</td>
<td>Old</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>AD or MCI</td>
<td>302</td>
</tr>
</tbody>
</table>

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
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, 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.

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 (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). 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. 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>
<thead>
<tr>
<th>Feature types</th>
<th>Model</th>
<th>AUC</th>
<th>F-score</th>
<th>F-score baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cov-ling-phon-</td>
<td>Svm</td>
<td>0.92</td>
<td>0.91</td>
<td>0.83</td>
</tr>
<tr>
<td>Cov-phon-pos-</td>
<td>Svm</td>
<td>0.92</td>
<td>0.97</td>
<td>0.91</td>
</tr>
<tr>
<td>Cov-ling-phon-pos-</td>
<td>Svm</td>
<td>0.90</td>
<td>0.93</td>
<td>0.81</td>
</tr>
<tr>
<td>Markers-ling-phon-pos-</td>
<td>Svm</td>
<td>0.89</td>
<td>0.96</td>
<td>0.77</td>
</tr>
<tr>
<td>Cov-phon-</td>
<td>Svm</td>
<td>0.89</td>
<td>0.90</td>
<td>0.91</td>
</tr>
<tr>
<td>Markers-ling-</td>
<td>Svm</td>
<td>0.88</td>
<td>0.89</td>
<td>0.85</td>
</tr>
<tr>
<td>Markers-cov-ling-</td>
<td>Svm</td>
<td>0.88</td>
<td>0.86</td>
<td>0.82</td>
</tr>
<tr>
<td>Markers-cov-ling-</td>
<td>Rfc</td>
<td>0.86</td>
<td>0.86</td>
<td>0.82</td>
</tr>
<tr>
<td>Markers-ling-phon-</td>
<td>Svm</td>
<td>0.86</td>
<td>0.90</td>
<td>0.81</td>
</tr>
<tr>
<td>Markers-cov-phon-pos-</td>
<td>Svm</td>
<td>0.86</td>
<td>0.93</td>
<td>0.86</td>
</tr>
<tr>
<td>Pitt Corpus (English)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Markers-cov-phon-pos-</td>
<td>Svm</td>
<td>0.76</td>
<td>0.79</td>
<td>0.68</td>
</tr>
<tr>
<td>Markers-cov-</td>
<td>Svm</td>
<td>0.74</td>
<td>0.77</td>
<td>0.67</td>
</tr>
<tr>
<td>Markers-cov-ling-pos-</td>
<td>Svm</td>
<td>0.74</td>
<td>0.77</td>
<td>0.69</td>
</tr>
<tr>
<td>Markers-cov-ling-</td>
<td>Svm</td>
<td>0.73</td>
<td>0.77</td>
<td>0.66</td>
</tr>
<tr>
<td>Markers-phon-pos-</td>
<td>Svm</td>
<td>0.73</td>
<td>0.76</td>
<td>0.69</td>
</tr>
<tr>
<td>Markers-cov-pos-</td>
<td>Svm</td>
<td>0.73</td>
<td>0.76</td>
<td>0.69</td>
</tr>
<tr>
<td>Markers-cov-phon-</td>
<td>Svm</td>
<td>0.73</td>
<td>0.75</td>
<td>0.68</td>
</tr>
<tr>
<td>Markers-ling-phon-pos-</td>
<td>Svm</td>
<td>0.73</td>
<td>0.75</td>
<td>0.70</td>
</tr>
<tr>
<td>Markers-ling-pos-</td>
<td>Svm</td>
<td>0.72</td>
<td>0.74</td>
<td>0.67</td>
</tr>
<tr>
<td>Markers-cov-ling-phon-</td>
<td>Svm</td>
<td>0.72</td>
<td>0.75</td>
<td>0.69</td>
</tr>
</tbody>
</table>

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 num-
ber 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 cor-
related 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 nega-
tive (-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 distri-
bution 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 consist-
tent 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 seg-
mented the audio to remove the interviewer’s voice. For the CRIUGM
data set, some of the MFCCs’ mean, skewness, and variance values
were highly correlated with the disease (> 0.4). Again, those high cor-
rrelations 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. 8 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 sus-
pect 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 (includ-
ing 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 sig-
nificantly distinguish AD patients from healthy controls. 7 Finally, we
believe it would be interesting to compare results between proportion-
ate 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
Ecole de technologie supérieure (H20170506).

CONFLICTS OF INTEREST

The authors have declared that no conflicts of interest exists.

REFERENCES

itudes to dementia. London, UK: Alzheimer’s Disease International;
2019.
developing dementia: predictive dementia screening from biographic
interviews. In: 2019 IEEE Automatic Speech Recognition and Understand-
3. You Y et al. Predicting dementia risk using paralinguistic and mem-
ory test features with machine learning models. In: 2019 IEEEHealth-

Howev


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