

LREC 2020  
Language Resources and Evaluation Conference  
11-16 May 2020

**3rd RaPID Workshop:  
Resources and Processing of Linguistic,  
Para-linguistic and Extra-linguistic Data from  
People with Various Forms of  
Cognitive/Psychiatric/Developmental  
Impairments**

**PROCEEDINGS**

Dimitrios Kokkinakis, Kristina Lundholm Fors,  
Charalambos Themistocleous, Malin Antonsson, Marie  
Eckerström (eds.)

**Proceedings of the LREC 2020 Workshop on:  
Resources and Processing of Linguistic, Para-linguistic  
and Extra-linguistic Data from People with Various Forms of  
Cognitive/Psychiatric/Developmental Impairments (RaPID-3)**

Edited by:

Dimitrios Kokkinakis, Kristina Lundholm Fors, Charalambos Themistocleous,  
Malin Antonsson, Marie Eckerström

**ISBN: 979-10-95546-45-0**

**EAN: 9791095546450**

**Acknowledgments:** This work has received support from the *Swedish Foundation for Humanities and Social Sciences* (RJ) through the grant agreement no: NHS14-1761:1 and the *Centre for Ageing and Health* (AgeCap, <https://agecap.gu.se/>).



**RIKSBANKENS  
JUBILEUMSFOND**  
STIFTELSEN FÖR HUMANISTISK OCH  
SAMHÄLLSVETENSKAPLIG FORSKNING



**For more information:**

European Language Resources Association (ELRA)

9 Rue des Cordelières

75013 Paris

France

<http://www.elra.info>

Email : [info@elda.org](mailto:info@elda.org)

© European Language Resources Association (ELRA)

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

## RaPID3@LREC2020 - Preface

Welcome to the LREC2020 Workshop on "Resources and Processing of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric/developmental impairments" (RaPID-3).

RaPID-3 aims to be an interdisciplinary forum for researchers to share information, findings, methods, models and experience on the collection and processing of data produced by people with various forms of mental, cognitive, neuropsychiatric, or neurodegenerative impairments, such as aphasia, dementia, autism, bipolar disorder, Parkinson's disease or schizophrenia. Particularly, the workshop's focus is on creation, processing and application of data resources from individuals at various stages of these impairments and with varying degrees of severity. Creation of resources includes e.g. annotation, description, analysis and interpretation of linguistic, paralinguistic and extra-linguistic data (such as spontaneous spoken language, transcripts, eyetracking measurements, wearable and sensor data, etc). Processing is done to identify, extract, correlate, evaluate and disseminate various linguistic or multimodal phenotypes and measurements, which then can be applied to aid diagnosis, monitor the progression or predict individuals at risk.

A central aim is to facilitate the study of the relationships among various levels of linguistic, paralinguistic and extra-linguistic observations (e.g., acoustic measures; phonological, syntactic and semantic features; eye tracking measurements; sensors, signs and multimodal signals). Submission of papers are invited in all of the aforementioned areas, particularly emphasizing multidisciplinary aspects of processing such data and the interplay between clinical/nursing/medical sciences, language technology, computational linguistics, natural language processing (NLP) and computer science. The workshop will act as a stimulus for the discussion of several ongoing research questions driving current and future research by bringing together researchers from various research communities.

### Topics of Interest

The topics of interest for the workshop session include but are not limited to:

- Infrastructure for the domain: building, adapting and availability of linguistic resources, data sets and tools
- Methods and protocols for data collection
- Acquisition and combination of novel data samples; including techniques for continuous streaming, monitoring and aggregation; as well as self-reported behavioral and/or physiological and activity data
- Guidelines, protocols, annotation schemas, annotation tools
- Addressing the challenges of representation, including dealing with data sparsity and dimensionality issues, feature combination from different sources and modalities
- Domain adaptation of NLP/AI tools
- Acoustic/phonetic/phonologic, syntactic, semantic, pragmatic and discourse analysis of data; including modeling of perception (e.g. eye-movement measures of reading) and production processes (e.g. recording of the writing process by means of digital pens, keystroke logging etc.); use of gestures accompanying speech and non-linguistic behavior

- Use of wearable, vision, and ambient sensors or their fusion for detection of cognitive disabilities or decline
- (Novel) Modeling and deep / machine learning approaches for early diagnostics, prediction, monitoring, classification etc. of various cognitive, psychiatric and/or developmental impairments
- Evaluation of the significance of features for screening and diagnostics
- Evaluation of tools, systems, components, metrics, applications and technologies including methodologies making use of NLP; e.g. for predicting clinical scores from (linguistic) features
- Digital platforms/technologies for cognitive assessment and brain training
- Evaluation, comparison and critical assessment of resources
- Involvement of medical/clinical professionals and patients
- Ethical and legal questions in research with human data in the domain, and how they can be handled
- Deployment, assessment platforms and services as well as innovative mining approaches that can be translated to practical/clinical applications
- Experiences, lessons learned and the future of NLP/AI in the area

## Submissions

Papers were invited in all of the areas outlined in the Topics of interest, particularly emphasizing multidisciplinary aspects of processing such data and the interplay between clinical/nursing/medical sciences, language technology, computational linguistics, NLP, and computer science. We welcomed also papers discussing problems derived from the design of relevant data samples and populations, but also the exploitation of results and outcomes as well as legal and ethical questions on how to deal with such data and make it available. Furthermore, the workshop solicited papers describing original research; and preferably describing substantial and completed work, but also focused on a contribution, a negative result, an interesting application nugget, a software package, a small, or work in progress. The workshop acted as a stimulus for the discussion of several ongoing research questions driving current and future research and challenges by bringing together researchers from various research communities.

***We are grateful to our Program Committee members for their hard work in reading and evaluating all submissions.*** At the end, each submission received between 2 to 5 reviews, which helped the authors revise and improve their papers accordingly.

Unfortunately the workshop, which was originally planned to take place on the 11th of May 2020 in conjunction with the LREC 2020 conference, could not be held as a face-to-face meeting due to the ongoing Covid-19 pandemic. Nevertheless, there were 18 contributions accepted for the workshop (6 to be oral presentations and 12 to be posters). A keynote talk was invited by Dr. Athanasios Tsanas, the Usher Institute, University of Edinburgh, UK, with the title: "Harnessing voice signals using signal processing and statistical machine learning: applications in mental health and other biomedical and life sciences applications".

Workshop website: <https://spraakbanken.gu.se/en/rapid-2020>.

**Organizers:**

Dimitrios Kokkinakis, University of Gothenburg, Sweden (*Workshop's chair*)  
Kristina Lundholm Fors, University of Gothenburg, Sweden  
Graeme Hirst, University of Toronto, Canada  
Malin Antonsson, University of Gothenburg, Sweden  
Charalambos Themistocleous, Johns Hopkins University, Baltimore, USA  
Marie Eckerström, University of Gothenburg, Sweden

**Program Committee (in alphabetic order):**

Jan Alexandersson, DFKI GmbH, Germany  
Malin Antonsson, University of Gothenburg, Sweden  
Eiji Aramaki, Nara Institute of Science and Technology (NAIST), Japan  
Visar Berisha, Arizona State University, USA  
Ellen Breitholtz, University of Gothenburg, Sweden  
Marie Eckerström, the Sahlgrenska Academy, University of Gothenburg, Sweden  
Valantis Fyndanis, University of Oslo, Norway  
Peter Garrard, St George's, University of London, UK  
Kallirroï Georgila, University of Southern California, USA  
Annette Gerstenberg, University of Potsdam, Germany  
Katarina Heimann Mühlenbock, University of Gothenburg, Sweden  
Graeme Hirst, University of Toronto, Canada  
Christine Howes, University of Gothenburg, Sweden  
Dimitrios Kokkinakis, University of Gothenburg, Sweden  
Alexandra König, Geriatric Hospital Nice and the University of Côte d'Azur, France  
Nicklas Linz, DFKI GmbH, Germany  
Peter Ljunglöf, University of Gothenburg, Sweden  
Kristina Lundholm Fors, University of Gothenburg, Sweden  
Saturnino Luz, University of Edinburgh, UK  
Juan José García Meilán, Universidad de Salamanca, Spain  
Mauro Nicolao, The University of Sheffield, UK  
Alexandre Nikolaev, Helsinki Collegium for Advanced Studies, Finland  
Marcus Nyström, University of Lund, Sweden  
Aurélië Pistono, Ghent University, Belgium  
Vassiliki Rentoumi, SKEL, NCSR Demokritos, Greece  
Fabien Ringeval, Université Grenoble Alpes, France  
Frank Rudzicz, Toronto Rehabilitation Institute and the University of Toronto, Canada  
Ineke Schuurman, KU Leuven, Belgium  
Kairit Sirts, University of Tartu, Estonia  
Charalambos Themistocleous, Johns Hopkins University, Baltimore, USA  
Athanasios Tsanas, the Usher Institute, University of Edinburgh, UK  
Magda Tsolaki, Aristotle University of Thessaloniki, Greece  
Spyridoula Varlokosta, National and Kapodistrian University of Athens, Greece  
Yasunori Yamada, IBM Research, Tokyo, Japan  
Stelios Zygoris, Aristotle University of Thessaloniki, Greece

**Invited Speaker:**

Athanasios Tsanas, the Usher Institute, University of Edinburgh, UK.

## Table of Contents

<i>Dependency Analysis of Spoken Language for Assessment of Neurological Disorders</i> Elif Eyigoz, Mary Pietrowicz, Carla Agurto, Juan Rafael Orozco, Adolfo M. Garcia, Sabine Skodda, Jan Ruzs, Elmar Nöth and Guillermo Cecchi.....	1
<i>Predicting Self-Reported Affect from Speech Acoustics and Language</i> Chelsea Chandler, Peter Foltz, Jian Cheng, Alex S. Cohen, Terje B. Holmlund and Brita Elvevåg.....	9
<i>The RiMotivAzione Dialogue Corpus - Analysing Medical Discourse to Model a Digital Physiotherapist</i> Francesca Alloatti, Andrea Bolioli, Alessio Bosca and Mariafrancesca Guadalupi.....	16
<i>Automatic Quantitative Prediction of Severity in Fluent Aphasia Using Sentence Representation Similarity</i> Katherine Ann Dunfield and Günter Neumann.....	24
<i>Linguistic Markers of Anorexia Nervosa: Preliminary Data from a Prospective Observational Study</i> Giulia Minori, Gloria Gagliardi, Vittoria Cuteri, Fabio Tamburini, Elisabetta Malaspina, Paola Gualandi, Francesca Rossi, Filomena Moscano, Valentina Francia and Antonia Parmeggiani.....	34
<i>What Difference Does it Make? Early Dementia Detection Using the Semantic and Phonemic Verbal Fluency Task</i> Hali Lindsay, Johannes Tröger, Jan Alexandersson and Alexander König.....	46
<i>Toward Characterizing the Language of Adults with Autism in Collaborative Discourse</i> Christine Yang, Emily Prud'hommeaux, Laura B. Silverman and Allison Canfield.....	54
<i>Automatic Classification of Primary Progressive Aphasia Patients Using Lexical and Acoustic Features</i> Sunghye Cho, Naomi Nevler, Sanjana Shellikeri, Sharon Ash, Mark Liberman and Murray Grossman.....	60
<i>Affective Speech for Alzheimer's Dementia Recognition</i> Fasih Haider, Sofia de la Fuente, Pierre Albert and Saturnino Luz.....	67
<i>Individual Mandibular Motor Actions Estimated from Speech Articulation Features</i> Andrés Gómez-Rodellar, Athanasios Tsanas, Pedro Gómez-Vilda, Agustín Álvarez-Marquina and Daniel Palacios-Alonso.....	74
<i>Digital Eavesdropper – Acoustic Speech Characteristics as Markers of Exacerbations in COPD Patients</i> Julia Merkus, Ferdy Hubers, Catia Cucchiarini and Helmer Strik.....	78
<i>Latent Feature Generation with Adversarial Learning for Aphasia Classification</i> Anna Vechkaeva and Günter Neumann.....	88
<i>Automated Analysis of Discourse Coherence in Schizophrenia: Approximation of Manual Measures</i> Galina Ryazanskaya and Mariya Khudyakova.....	98
<i>The Mind-It Corpus: a Longitudinal Corpus of Electronic Messages Written by Older Adults with Incipient Alzheimer's Disease and Clinically Normal Volunteers</i> Olga Semink, Louise-Amélie Coughn, Bernard Hanseeuw and Cédric Fairon.....	108

<i>Coreference in Aphasic and non-Aphasic Spoken Discourse: Annotation Scheme and Preliminary Results</i>	
Svetlana Toldova, Elizaveta Ivtushok, Kira Shulgina and Mariya Khudyakova.....	116
<i>An NLP pipeline as assisted transcription tool for speech therapists</i>	
Gloria Gagliardi, Lorenzo Gregori and Andrea Amelio Ravelli.....	124
<i>An Exploration of Personality Traits Detection in a Spanish Twitter Corpus</i>	
Gerardo Sierra, Gemma Bel-Enguix, Alejandro Osornio-Arteaga, Adriana Cabrera-Mora, Luis García-Nieto, Alfredo Bustos, Ana-Miriam Romo-Anaya and Víctor Silva-Cuevas.....	132
<i>Using Dependency Syntax-Based Methods for Automatic Detection of Psychiatric Comorbidities</i>	
Yannis Haralambous, Christophe Lemey, Philippe Lenca, Romain Billot and Deok-Hee Kim-Dufor.....	142



# Dependency Analysis of Spoken Language for Assessment of Neurological Disorders

Elif Eyigoz<sup>1</sup>, Mary Pietrowicz<sup>1</sup>, Carla Agurto<sup>1</sup>  
Juan Rafael Orozco<sup>2</sup>, Adolfo M. Garcia<sup>3</sup>, Sabine Skodda<sup>4</sup>  
Jan Rusz<sup>5</sup>, Elmar Nöth<sup>6</sup>, Guillermo Cecchi<sup>1</sup>

<sup>1</sup>IBM Research USA, ekeyigoz, gcecchi@us.ibm.com, mary.pietrowicz, carla.agurto@ibm.com

<sup>2</sup>University of Antioquia Colombia, adolfomartingarcia@gmail.com

<sup>3</sup>Favaloro University Argentina, adolfomartingarcia@gmail.com

<sup>4</sup>University of Bochum Germany, sabine.skodda@kk-bochum.de

<sup>5</sup>Czech Technical University in Prague Czech Republic, ruszjan@fel.cvut.cz

<sup>6</sup>Friedrich-Alexander University of Erlangen Germany, elmar.noeth@fau.de

## Abstract

Spoken language carries a greater amount of information about the speakers' cognitive state compared to written language, when analysis of disfluencies and expressivity is considered. For the same reason, however, spoken language presents challenges for automated syntactic analysis. This study presents comparative performance of different preprocessing methods applied to spoken language data for syntactic analysis. Furthermore, this study presents a novel dependency tree analysis for assessment of neurological disorders. The methods described in this paper were compared against a baseline model consisting of features typically considered to provide for the largest amount of information within the context of neurological disorders. Significant improvements were obtained across multiple languages and multiple neurological disorders: The method improved detection of Alzheimer's disease patients in English-speaking subjects and detection of Parkinson's disease patients in German-speaking subjects.

**Keywords:** Assessment of neurological disorders, Spoken language processing, Syntactic analysis, Dependency grammar, Alzheimer's disease, Parkinson's disease

## 1. Introduction

Syntactic production and comprehension deficits have been well documented in a large variety of cognitive and neurological disorders, such as Parkinson's disease and Alzheimer's disease (Lyons et al., 1994; Lieberman et al., 1992; Bocanegra et al., 2015). As a result, the number of studies in the last decade that include, to a greater or lesser extent, automated syntactic analysis for assessment of neurological disorders has surged (Fraser et al., 2015; Orimaye et al., 2017; Aluísio et al., 2016; König et al., 2018; Eyigöz et al., 2018a; Roark et al., 2007; Eyigoz et al., 2018b). These studies hold out the promise that fully automated screening of individuals for neurological disorders may be within reach.

It is a reasonable claim that an automatic speech recognition (ASR) based system for assessment of neurological disorders that only uses textual analysis might not work as well as one which has the audio available for acoustic analysis. Nevertheless, it is still plausible to have systems with one of the design goals as compatibility with ASR, due to privacy and data sharing restrictions.

Prior methods for automated syntactic analysis for assessment of neurological disorders can be broadly classified into two categories: language independent methods (Eyigöz et al., 2018a; Roark et al., 2007), and language dependent methods (Fraser et al., 2015; Orimaye et al., 2017). Language dependent methods assume prior linguistic knowledge of node or edge labels on syntax trees for feature extraction, as they focus on a subset of edge or node labels (e.g noun phrases, verb phrases, prepositions, subordinate clauses etc), which are language specific. In order to apply these methods to, for example Japanese syntax

trees, one would need to find the counterparts of specific node/edge labels in the Japanese grammar. This would require an intensive statistical and/or linguistic effort. Language independent methods, on the other hand, make no assumption of prior knowledge of tree labels. The claim of "language-independence" does not mean to imply that these methods can be applied without a language-specific syntactic parser. Instead, it means that if a syntactic parser exists for a given language, then the method can be applied *without any modification* to this new language as well. This would not be possible if the method assumes certain node/edge labels on syntax trees.

Spoken language carries great amount of information on the speakers' cognitive state, as it is amenable to analysis through observation of speech disfluencies, such as spontaneous errors, fillers, pauses, false starts, etc. For the same reason, however, spoken language poses a challenge for automated syntactic analysis. Syntactic parsers are traditionally trained on written language, thus are prone to making errors on spoken language data. Furthermore, language of subjects with neurological disorders exhibit even higher levels of speech disfluencies. Consequently, special care has to be taken in preprocessing of spoken language data produced by patients of neurological disorders before syntactic parsing.

This study presents a language-independent dependency-tree based syntactic analysis method, which is agnostic of the edge labels on the dependency trees. Furthermore, this study presents comparative performance of different preprocessing methods applied to two different transcription styles for syntactic analysis. We compared the performance of our method to a baseline model consisting of features typically considered to provide for the largest amount of

information within the context of neurological disorders. We obtained significant improvements over the baseline method in classification of Alzheimer’s disease patients in English-speaking subjects, and classification of Parkinson’s disease patients in German-speaking subjects. Therefore, our experiments suggest that the proposed method may generalize across different languages and neurological conditions.

## 2. Preprocessing of spoken language for syntactic parsing in two datasets

### 2.1. Alzheimer’s disease

Alzheimer’s disease (AD) is an aging related neurodegenerative disorder that progressively destroys memory and cognitive skills. For the AD experiments, we used the publicly available DementiaBank dataset (Forbes et al., 2012), which has been used in more than 100 research studies. Given the popularity of the DementiaBank dataset, we believe a comparison of preprocessing methods of this dataset would be of interest to a large audience.

For our study, we used 257 speech samples from 169 Alzheimer’s disease patients, and 242 samples from 99 controls in English from the DementiaBank dataset. The subjects were asked to describe the Cookie-theft picture (Goodglass et al., 2000), and the recordings were transcribed with the CHAT coding convention (MacWhinney, 2000). CHAT was designed for transcription of spoken language, thus allows for detailed annotation of a large variety of speech phenomena. Spoken language data annotated in such level of detail is not widely available. Instead, systems designed for fully automated analysis of text for screening of individuals will have access to the outputs of ASR systems. We experimented with two different preprocessing methods of the CHAT files, both of which would be plausible in the ASR setting. For the first one, we replicated the preprocessing method in (Fraser et al., 2015), which was proposed for CHAT files. We refer to this preprocessing method as the `light` preprocessing method. This method left all disfluencies (repetitions, revisions, paraphasias) in the transcript, removed all short false starts of two letters or fewer (*c-cookie*, *th-theft*), all filled pauses such as (*um*, *uh*, *er*), and all other meta annotation.

In addition, we implemented a second preprocessing method (`heavy`) that is still plausible in the ASR setting, and is aimed at increasing syntactic parsing accuracy. We removed all phonological fragments shorter than two characters (which includes all false starts of less than two characters), removed all word and phrase repetitions, replaced the nonstandard forms such as assimilations (*lemme* ⇒ *let me*; *gonna* ⇒ *going to*), shortenings (*bout* ⇒ *about*; *cause* ⇒ *because*), phonological variations (*deir* ⇒ *their*; *ya* ⇒ *you*) with their standard forms. As for insertion of punctuation, we replaced all filled pauses (*um*, *uh*, *er*) with a comma, replaced short pauses with a space, replaced long pauses with a comma, replaced very long pauses with a full stop. We put a comma after phrase revisions and reformulations. We did not insert a comma or a full stop if the position of insertion comes right after a determiner, a preposition or a copula, because we assumed that an automated punctuation insertion model trained on written language

would not insert punctuation at these positions. Finally, we obtained the dependency parse trees using the Stanford CoreNLP parser.

### 2.2. Parkinson’s disease

Parkinson’s disease (PD) is a neurodegenerative disorder that causes gradual loss of muscle control, which results in muscle rigidity, tremors, and changes in gait and speech. The PD dataset used in this study consists of a unique sample in German from 88 PD patients and 88 cognitively normal controls, who were asked to describe a typical day in their lives (Skodda et al., 2011). A single person transcribed the data, and employed a simplistic transcription scheme, which resulted in a dataset more similar to ASR output than CHAT transcriptions.

The transcriber used ‘...’ to annotate all disfluencies, long pauses, and after fillers, as well as the end of fragmented utterances. We implemented two preprocessing methods for the German data: In the `naive` preprocessing scheme, we removed all meta annotation (e.g. [laughing]), and then we replaced all ‘...’ with a space. The alternative `light` preprocessing scheme required an examination of the transcriptions for discovering the regularities in transcription style and punctuation use, and for discovering the filler words. The filler words were determined as highly frequent words that were out-of-vocabulary, and they were removed along with the ‘...’ after them. If there were no punctuation before a line break, a full stop was inserted. If a line ended with ‘...’, it was replaced with a full stop.

The inspection of the data revealed that replacing all the remaining ‘...’ with a space would concatenate numerous utterances into one large utterance with no utterance boundaries. On the other hand, replacing ‘...’ with a full stop would split grammatical utterances into ungrammatical fragments. Therefore, we decided to indicate any disfluency annotated by a ‘...’ to the syntactic parser with a comma. For repetitions of more than two words, only the first two were kept, in order not to lose signal, but to eliminate large parsing errors.

Finally, we obtained the dependency parse trees using the Stanford CoreNLP parser.

## 3. Dependency tree analysis

Figure 1 shows a dependency parse tree, which depicts dependency relations between word pairs with an outgoing arrow. In the relation between *this* and *paper*, the arrow originates from *paper* and points to *this*. In dependency trees, an arrow originates from the head word, and points to the dependent word, therefore *paper* is the head, and *this* is the dependent in this particular relation. In dependency trees, relations have types as indicated by the labels on the arrows, e.g. in Figure 1, the relation between *this* and *paper* is a determiner (*det*) relation. The number of relations (edges) in a dependency tree equals the number of words, which in this example is 11.

Rates or counts of dependency-relation types and part-of-speech tags (pos-tags) have been widely used for analysis of syntactic complexity (Fraser et al., 2015; Orimaye et al., 2014; Rudzicz et al., 2014). For example, there are two *det* relations in Figure 1 (arrows

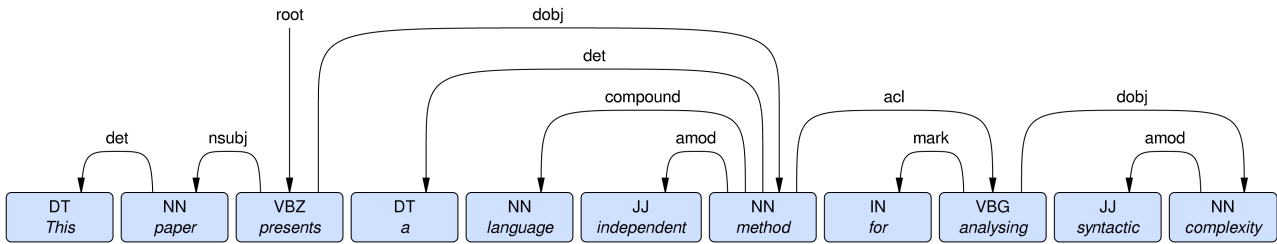


Figure 1: Example dependency parse tree.

pointing to *this* and *a*), therefore the rate of *det* relation per word is 2/11. Likewise, we computed the rates of dependency relations per utterance, i.e. normalizing the counts of dependency relations by the total number of utterances, and rates of pos-tags, normalized by either the number of words or the number of utterances. We refer to these features as `relation-per-word`, `pos-per-word`, `relation-per-utterance`, `pos-per-utterance` respectively in Section 6.

The out-degree of a node is the number of arrows originating from the node; for example, the out-degree of node *method* is 4 in Figure 1. The depth of a dependency tree is the length of the longest path from its root to one of its leaves. We used standard deviation, median, percentile 10 and 90, skewness and kurtosis of the depth of the trees and the out-degrees of nodes as features. Features that analyze the depth of syntax trees (Yngve, 1960; Sampson, 1997), and features using statistics over the length of context-free-grammar rules (Fraser et al., 2015), which are conceptually similar to the out-degrees of the nodes in dependency trees, were suggested previously as measures of syntactic complexity.

Finally, we developed a novel syntactic feature extraction method that analyzes larger subparts of the parse tree than a single dependency relation. For this purpose, we considered relation-pairs that are consecutive in the dependency graph. Two arrows of a directed graph are consecutive if the arrow of the first one is at the nock of the second one. For example in Figure 1, the relations *nsubj* and *det* are consecutive in *this paper presents*. Please note that the consecutive relation is not symmetric, e.g. *nsubj* and *det* are consecutive, but *det* and *nsubj* are not consecutive in *this paper presents*. Similarly, *root* and *nsubj* are consecutive in *paper presents*. In Figure 1, some consecutive relation pairs occur more than once: for example the *dobj* and *amod* pair occurs twice: first in *independent method*, second in *syntactic complexity*. For each instance of a consecutive relation pair, we count the number of their occurrences, and normalize them by the total number words or by the total number utterances in the sample. For example, the rate of `consecutive(dobj,amod)` per words is 2/11 in Figure 1. We refer to these features as `consecutive-per-word`, `consecutive-per-utterance` respectively in Section 6. Figure 1 shows the parse tree of a single sentence, however speech samples usually consist of multi-

ple utterances. Consequently, we collected counts for each dependency-relation type, each consecutive relation pair, and each pos-tag across multiple utterances in a given sample.

#### 4. Baseline model for neurological disorders

As a baseline model, we used features that were frequently shown to carry the strongest signal in studies for assessment of neurological disorders including verbosity, lexical diversity/richness, repetitiveness, and word-frequencies. The verbosity features were total number of words, total number of utterances, total number of types, total number of characters, and characters per word. Lexical diversity/richness features were type-token ratio, Dugast (Dugast, 1979) and Guiraud (Guiraud, 1954) metrics. The word-frequency features were standard deviation, median, percentile 10 and 90, and kurtosis of word frequencies; and perplexity was computed as the mean of the logarithm of the word frequencies. Finally, the repetitiveness metric as defined in (Fraser et al., 2015) was computed on the lemmatized data, and the mean and the standard deviation of this metric was used as a feature.

#### 5. Experimental Method

Because our method does not assume prior knowledge of what dependency relations or pos-tags indicate in terms of syntactic complexity, we created a feature for all observed pos-tags (e.g. *NN*), for all observed dependency-relation types (e.g. *det*), and for all observed consecutive relation pairs (e.g. `consecutive(nsubj, det)`) in the entire dataset. As a result, this method generated a large number of features. For feature selection, we used the Wilcoxon rank-sum test (Wilcoxon et al., 1970) to compute a p-value for each feature, and eliminated features that were not statistically significant (p-value < .05). We used the Wilcoxon rank-sum test, because the features did not follow a normal distribution. We applied this feature selection method within a *k*-fold cross-validation setting (CV) as follows: We split the data into folds of train-test sets, such that data from any individual subject occurred in the test set or the training set, but not in both. Feature selection was applied within each CV fold using only the training data available to that fold, not observing the left-out samples for feature selection. The selected features were then used for training, and the labels of the left-out samples were predicted using the fitted classifiers.

Table 1: Features used in experiments in Table 2

Experimental setting	Features
Baseline	verbosity, lexical diversity/richness, repetitiveness, word-frequencies
Depth + Out	standard deviation, median, percentile 10 and 90, skewness and kurtosis of the depth of the trees and the out-degrees of nodes
Relation	relation-per-word
Pos-tag	pos-per-utterance
Consecutive	consecutive-per-utterance

We experimented with the following classifiers, and for each experimental setting, we report the results of the classifier with the highest accuracy score in the following section: Linear SVM, RBF SVM, Logistic regression, Naive-Bayes, Decision Trees.

## 6. Results

We performed 10-fold CV experiments for the English-AD dataset and 25-fold CV experiments for the German-PD dataset, since the English dataset was more than twice the size of the German dataset. We repeated the CV experiments 50 times, which enabled us to perform  $t$ -tests between the results of the settings to examine whether the differences were statistically significant. Table 2 shows the results of the experiments with their means across the 50 runs.

In Table 2, SS stands for statistical significance, and if the value is =, then the difference over/from the baseline is not statistically significant. If the value in Table 2 is < or >, then the  $p$ -value of the  $t$ -test between the results of the experimental setting and the baseline model is between .05 and .0001. In other words, < and > stand for comparisons with the baseline model where the improvement (>) or the deterioration (<) over the baseline model is statistically significant, but it is not a very strong difference (not  $p$ -val < .0001). For stronger results, we used the  $\gg$  or the  $\ll$  notation. If the value in Table 2 is  $\gg$  or  $\ll$ , then the  $p$ -value of the  $t$ -test between the results of the experimental setting and the baseline model is < .0001.

The score pairs that share the same subscript in Table 2 show that the difference between the results of the experimental settings is statistically significant, ( $p=.003$  in pair 1;  $p<.0001$  in pairs 2, 3, 4). The results reported in Table 2 were obtained by linear SVM classifiers except for the German naive setting, which was obtained by a Naive-Bayes classifier.

The results of feature selection can be seen in Table 2, as well. The median number of features that were selected across the CV folds was computed for each CV experiment, which in turn were averaged across the 50 runs of the CV experiments. The averaged medians are shown in Table 2 in the #Feat column. The total number of all features, baseline and syntactic combined, are given in the first row next to the preprocessing name. These results show that the Wilcoxon rank-sum test could successfully be used as a feature selection method in all experimental settings.

In English and German experiments, we observed that

different normalization methods (normalization by number of words or by number of utterances) showed consistent improvements for certain features in both languages. We decided to focus the presentation of our results in Table 2 on features that showed consistency across English and German with respect to normalization. First, in both English and German experiments, we observed that pos-per-utterance performed better than pos-per-word, although both provided statistically significant improvements over the baseline. For dependency-relation features, however, we observed the opposite pattern: relation-per-word performed better than relation-per-utterance, again for both languages. Therefore, features obtained with the inferior normalization method in both languages were not included in the experimental settings in Table 2, as it is reflected in Table 1, which shows the features used in experimental settings in Table 2. Consequently, in the rest of the paper, the pos-tag features refer to pos-per-utterance and the dependency-relation features refer to relation-per-word.

Contrarily, consecutive-per-utterance and consecutive-per-word did not show a consistent cross-language or inter-language direction in terms of performance. We decided to exclude consecutive-per-word from the experimental settings in Table 2, because consecutive-per-word is relatively similar to relation-per-word (included in Table 2), since both are functions of counts of dependency relations, and both are normalized by the number of words. Consequently, in the rest of the paper, the consecutive-relation features refer to consecutive-per-utterance only. Again, Table 1 reflects this elimination.

### 6.1. Correlated syntactic features

We observed that some of the syntactic features computed with our method were highly correlated. We investigated the effect of correlations between the syntactic features using the English-AD dataset. This dataset was chosen for this analysis because the largest improvements were obtained with this dataset.

Since dependency-relation features are more informative of the syntactic structure than the pos-tags, we eliminated all pos-tag features that were correlated with a dependency-relation feature. The total number of pos-tag features with the heavy preprocessing method were 33, and 15 of them

Table 2: Results table: The total number of features (before feature selection) can be found in the headers next to the preprocessing method name. B stands for baseline, Acc stands for accuracy score, and #Feat stands for the number of features used in the experiments. SS stands for statistical significance. If the value of SS is =, then the difference over/from the baseline is not statistically significant. If the value is < or >, then the p-value is between .05 and .0001. If the value is >> or <<, then the p-value is < .0001.

	Accuracy	Accuracy-SS	AUC	AUC-SS	#Features
<b>English</b>					
<b>Heavy</b>					992
Baseline	0.72	=	0.80	=	12
B+Depth+Out	0.73	>>	0.81	>>	16
B+Relation	0.77	>>	0.85	>>	26
B+Pos-tag	0.77	>>	0.85	>>	26
B+Consecutive	0.76	>>	0.84	>>	30
B+ALL	0.79 <sub>1</sub>	>>	<b>0.87<sub>2</sub></b>	>>	61
<b>Light</b>					915
Baseline	0.72	=	0.79	=	11
B+Depth+Out	0.71	<	0.78	<<	15
B+Relation	0.76	>>	0.83	>>	22
B+Pos-tag	0.77	>>	0.84	>>	24
B+Consecutive	0.77	>>	0.84	>>	31
B+ALL	<b>0.80<sub>1</sub></b>	>>	0.86 <sub>2</sub>	>>	57
<b>German</b>					
<b>Naive</b>					998
Baseline	0.62	=	0.63	=	5
B+Depth+Out	0.63	>	0.67	>>	7
B+Relation	0.61	<<	0.62	<	6
B+Pos-tag	0.64	>>	0.69	>>	9
B+Consecutive	0.61	<	0.62	=	12
B+ALL	0.62	=	0.66	>>	20
<b>Light</b>					1030
Baseline	0.62	=	0.68	=	7
B+Depth+Out	0.65	>>	0.70	>>	15
B+Relation	0.65	>>	0.69	>>	9
B+Pos-tag	0.67	>>	0.70	>>	22
B+Consecutive	0.67 <sub>3</sub>	>>	<b>0.73<sub>4</sub></b>	>>	29
B+ALL	<b>0.68<sub>3</sub></b>	>>	0.72 <sub>4</sub>	>>	52

were eliminated due to a correlation with a dependency-relation feature. The correlated features were determined by a Kendall  $\tau$  correlation above 0.5 or below -0.5. 0.5 was chosen as a threshold, because information conveyed by pos-tags and dependency relations presumably overlap to a significant degree, and our aim was to eliminate redundancy as much as possible with a moderate threshold. We used Kendall  $\tau$  correlation instead of Pearson  $r$  correlation, because the features were not normally distributed.

Next, we eliminated consecutive-relation features that were correlated with either a dependency-relation feature, or with a pos-tag feature that was not eliminated in the previous step. The correlated features were determined by a Kendall  $\tau$  correlation above 0.7 or below -0.7. We chose a higher threshold for elimination of consecutive-relation features than pos-tag features, because we believe that the information conveyed by the consecutive-relation features cannot be reduced to the information conveyed by

the dependency-relation features. Despite this high threshold, this method eliminated 68 percent of the consecutive-relation features, which were 401 in total with the *heavy* preprocessing method.

After elimination of features as described above, we performed the CV experiments as explained in the previous section, and obtained exactly the same performances reported in Table 2.

## 7. Discussion

The chance level accuracy of the English-AD dataset is 0.51, and the German PD dataset is 0.5. The baseline model, including features for verbosity, lexical diversity/richness, repetitiveness, and word-frequencies, obtained a very significant improvement over the chance level accuracy and AUC. Any additional improvement obtained over such a strong baseline is thus due to purely syntactic complexity of the speakers' language.

In the English-AD experiments, we obtained eight points improvement in accuracy score, and seven points improvement in AUC score over the baseline model. In the German-PD experiments, we obtained six points improvement in accuracy score, and five points improvement in AUC score over the baseline method. The improvements over the baseline were greater for the English-AD dataset than the German-PD dataset. The main reason for this difference, we believe, is that AD is primarily a cognitive impairment, whereas PD is primarily a motor impairment. Besides, the English and the German data differed significantly in transcription quality.

A direct comparison of our results with state-of-the-art is not possible, due to our method of reporting results as averages of multiple runs of CV experiments and applying feature selection within the CV folds -as opposed to before CV experiments. Furthermore, the state-of-the-art results on the same English dataset include semantic and acoustic features. The focus of our study was analysis of preprocessing methods before syntactic analysis, and cross-linguistic investigation of syntactic features, and we believe that we covered a wide range of features that have been proposed in the literature for syntactic analysis for degenerative disorders. The fact that we did not include semantic and acoustic features could be considered a weakness of this study, as it makes it impossible to provide a direct comparison with the state-of-the-art results on the English dataset. However, we can still provide a comparison of the final accuracy scores. An accuracy score of 0.81 was reported on the DementiaBank dataset using features overlapping to a large extent with our baseline method, and syntactic analysis of constituency trees (Fraser et al., 2015). They additionally used acoustic and semantic features. Our method obtained a mean accuracy score of 0.8, however the maximum accuracy score obtained across 50 runs of the CV experiments was 0.81. Therefore, we obtained the same accuracy score without using acoustic and semantic features. An F1 score of 0.74 was reported on the DementiaBank dataset using again a large variety of features including syntactic and lexical analyses (Orimaye et al., 2014). We obtained 0.79 F-1 score on average, and a maximum F-1 score of 0.81. Therefore, our results are comparable to the studies using lexical and syntactic features on the same data set.

Our results show that the preprocessing method for syntactic analysis makes a significant difference in performance. In the German *naive* preprocessing experiments, using the wrong preprocessing method caused a drop in performance below the baseline in the experiments that used the dependency features, namely the dependency-relation and consecutive-relation features. However, the pos-tag features were less effected by the inaccurate parses caused by the *naive* method. The reason is that pos-tagging is a preprocessing step for syntactic parsing, not an outcome of it. Nevertheless, all scores were lower for the *naive* preprocessing method than the *light* preprocessing method in German-PD experiments.

Furthermore, our results showed that using a preprocessing method aimed at increasing syntactic parse accuracy may be worth the effort. First, the only experimental condition of the English-AD data set that performed worse than

the baseline occurred as a result of the *light* preprocessing method. The addition of the *depth+out-degree* features to the baseline caused a statistically significant drop in performance. This implies that the inaccurate parse trees obtained by this method differed from the accurate parse trees strongly with respect to their depth and the degree of branching at their nodes.

Second, the consecutive-relation features were the only exception in the comparison between the preprocessing methods for English. For all other individual feature groups, the scores obtained with the *heavy* preprocessing method were either greater than or equal to the *light* preprocessing method. Even the baseline model benefited from using a heavier preprocessing method. Again, except for the consecutive-relation features, the number of features that passed the p-value threshold was higher for the *heavy* preprocessing method than the *light* preprocessing method. Despite the fact that all other feature groups performed better with the *heavy* preprocessing method, combining the consecutive-relation features with all other features increased the accuracy score of the *light* preprocessing method over the *heavy* preprocessing method, proving the strength of the consecutive-relation features in this setting.

The pos-tag features performed better or the same as the dependency-relation features across all experiments. Considering only the syntactically viable preprocessing settings (excluding the German *naive* preprocessing setting), we observe that the pos-tag features perform better than the consecutive-relation features in only one setting (English-*heavy*), the consecutive-relation features perform better than the pos-tag features in only one setting (German-*light*), and they perform the same in one setting (English-*light*). The fact that the pos-tag features perform better than the consecutive-relation features in English-*heavy* but not in English-*light* may mean that more accurate pos-tagging with the *heavy* method eliminated the need for additional information conveyed by the consecutive-relation features, while this was not the case with the English-*light* method. The fact that the consecutive-relation features perform better than the pos-tag features in German-*light* setting may be due to the characteristics of the production deficiencies in PD.

An examination of the features that have survived the feature-selection process in both languages showed that our method made use of features that have commonly been suggested as relating to human-language processing capabilities, and have been used in prior work. These features fall in four categories: (1) tree patterns involving optional modifiers, e.g. adjectives or adverbs; (2) tree patterns involving sentence embeddedness; e.g. subordinate clauses; (3) tree patterns involving telegraphic speech, a typical symptom of aphasia, e.g. use of gerunds; and (4) tree patterns involving parsing errors due to fragmented utterances, and repetitions.

Unlike consecutive-relation features, the pos-tag features and dependency-relation features have been used in prior work. Nevertheless, our work for the first time investigated the shared characteristics of different normalization methods (per word or per node) across multiple languages for these frequently used features.

## 8. Conclusion

To summarize, the contributions of this study are (1) a comparison of preprocessing methods applied to different transcription styles for spoken language data; (2) the interaction of these preprocessing methods with previously suggested and novel syntactic features; (3) cross-linguistic investigation of previously suggested and novel syntactic features on multiple neurological disorders.

- Aluísio, S., Cunha, A., and Scarton, C. (2016). Evaluating progression of Alzheimer’s disease by regression and classification methods in a narrative language test in portuguese. In *Lecture Notes in Computer Science*, pages 109–114. Springer International Publishing.
- Bocanegra, Y., García, A. M., Pineda, D., Buriticá, O., Villegas, A., Lopera, F., Gómez, D., Gómez-Arias, C., Cardona, J. F., Trujillo, N., et al. (2015). Syntax, action verbs, action semantics, and object semantics in parkinson’s disease: Dissociability, progression, and executive influences. *Cortex*, 69:237–254.
- Dugast, D. (1979). *Vocabulaire et stylistique*, volume 8. Ed. Slatkine.
- Eyigöz, E., Cecchi, G., and Tejwani, R. (2018a). Predicting cognitive impairments with a mobile application. In *Proceedings of the 10th International Conference on Agents and Artificial Intelligence*. SCITEPRESS - Science and Technology Publications.
- Eyigoz, E., Polosecki, P., Garcia, A. M., Rogg, K., Orozco-Arroyave, J., Skodda, S., Hesse, E., Ibanez, A., and Cecchi, G. (2018b). Unsupervised morphological segmentation for detecting parkinson’s disease. In *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*.
- Forbes, M. M., Fromm, D., and MacWhinney, B. (2012). Aphasiabank: A resource for clinicians. In *Seminars in speech and language*, volume 33, pages 217–222. Thieme Medical Publishers.
- Fraser, K. C., Meltzer, J. A., and Rudzicz, F. (2015). Linguistic features identify Alzheimer’s disease in narrative speech. *Journal of Alzheimer’s Disease*, 49(2):407–422, Oct.
- Goodglass, H., Kaplan, E., and Barresi, B. (2000). *Boston Diagnostic Aphasia Examination Record Booklet*. Lippincott Williams & Wilkins.
- Guiraud, P. (1954). *Les caractères statistiques du vocabulaire*. Presses universitaires de France.
- König, A., Linz, N., Tröger, J., Rainouard, A., Gros, A., Alexandersson, J., and Robert, P. (2018). Automatic speech analysis for the detection of emotional disturbances in persons with dementia. *Alzheimer’s & Dementia: The Journal of the Alzheimer’s Association*, 14(7):P1465.
- Lieberman, P., Kako, E., Friedman, J., Tajchman, G., Feldman, L. S., and Jiminez, E. B. (1992). Speech production, syntax comprehension, and cognitive deficits in parkinson’s disease. *Brain and language*, 43(2):169–189.
- Lyons, K., Kemper, S., Labarge, E., Ferraro, F. R., Balota, D., and Storandt, M. (1994). Oral language and Alzheimer’s disease: A reduction in syntactic complexity. *Aging, Neuropsychology, and Cognition*, 1:271–281, Dec.
- MacWhinney, B. (2000). The chldes project: Tools for analyzing talk: Volume i: Transcription format and programs, volume ii: The database.
- Orimaye, S. O., Wong, J. S.-M., and Golden, K. J. (2014). Learning predictive linguistic features for Alzheimer’s disease and related dementias using verbal utterances. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Association for Computational Linguistics.
- Orimaye, S. O., Wong, J. S.-M., 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), jan.
- Roark, B., Mitchell, M., and Hollingshead, K. (2007). Syntactic complexity measures for detecting mild cognitive impairment. In *Proceedings of the Workshop on BioNLP 2007 Biological, Translational, and Clinical Language Processing - BioNLP07*. Association for Computational Linguistics.
- Rudzicz, F., Currie, L. C., Danks, A., Mehta, T., and Zhao, S. (2014). Automatically identifying trouble-indicating speech behaviors in Alzheimer’s disease. In *Proceedings of the 16th international ACM SIGACCESS conference on Computers & accessibility - ASSETS14*. ACM Press.
- Sampson, G. (1997). Depth in English grammar. *Journal of Linguistics*, 33(1):131–151.
- Skodda, S., Visser, W., and Schlegel, U. (2011). Vowel articulation in Parkinson’s disease. *Journal of Voice*, 25(4):467–472.
- Wilcoxon, F., Katti, S., and Wilcox, R. A. (1970). Critical values and probability levels for the wilcoxon rank sum test and the wilcoxon signed rank test. *Selected tables in mathematical statistics*, 1:171–259.
- Yngve, V. H. (1960). A model and an hypothesis for language structure. *Proceedings of the American philosophical society*, 104(5):444–466.



# Predicting Self-Reported Affect from Speech Acoustics and Language

Chelsea Chandler<sup>1,2</sup>, Peter W. Foltz<sup>2,3</sup>, Jian Cheng<sup>4</sup>, Alex S. Cohen<sup>5</sup>, Terje B. Holmlund<sup>6</sup>, Brita Elvevåg<sup>6,7</sup>

<sup>1</sup>Department of Computer Science, University of Colorado Boulder, CO, U.S.A.

<sup>2</sup>Institute of Cognitive Science, University of Colorado Boulder, CO, U.S.A.

<sup>3</sup>Pearson, CO, U.S.A. <sup>4</sup>Analytic Measures Inc., Palo Alto, CA, U.S.A.

<sup>5</sup>Department of Psychology, Louisiana State University, LA, U.S.A.

<sup>6</sup>Department of Clinical Medicine, UiT The Arctic University of Norway, Norway

<sup>7</sup>Norwegian Centre for eHealth Research, University Hospital of North Norway, Tromsø, Norway  
{chelsea.chandler, peter.foltz}@colorado.edu, brita@elvevaag.net

## Abstract

Communication via speech offers a window into a person’s mental and cognitive states. Both the manner in which a person speaks (acoustics) and the words spoken (language) may be used to assay current mental and cognitive function. In this study, we predicted self-reported emotion from the acoustics and language of the seemingly affectless task of verbally recalling a short story. Story recalls and self-reports of affect were collected over multiple days via a mobile application in a population of 21 psychiatric patients and 79 presumed healthy participants, resulting in 137 and 430 total sessions for each group respectively. We have previously shown that analyzing just one modality of data produces moderate correlations with self-reported affect ( $0.33 < r < 0.40$  for speech and  $0.07 < r < 0.28$  for language). The goal of this study was to improve on unimodal analyses by extracting acoustic and language features from story recalls and combining them to predict a person’s self-reported affect. This combination of modalities resulted in an improvement over just one modality alone ( $0.38 < r < 0.48$ ). We show that a multimodal analytic approach predicted self-reported emotional states in clinical and non-clinical participants better than a unimodal approach.

**Keywords:** positive affect, negative affect, speech, natural language processing, machine learning

## 1. Introduction

Psychosis can disrupt language production in a number of different ways, including acoustics (transmitted sound), lexemes (word choice), syntax (sentence structure), coherence (logical flow), and semantics (meaning; see Holmlund et al., 2020b for a review). Therefore, the evaluation of language production is an important component during clinical interviews and in many standard psychosis rating scales. Such formal examinations would hugely benefit from more objective and rigorously defined analyses where automation can improve speed, reliability and consistency of judgments. In this study, we first sought to answer whether machine learning techniques could infer what aspects of speech acoustics and what words and patterns of words in language could be used to predict positive and negative affect, and second whether the combination of these two modalities would improve on unimodal predictions.

Clinically valuable characteristics, notably negative affect, exist even just in the sound of patients’ voices (Cohen et al., 2016a, 2016b). An example of such a characteristic is a lack of vocal modulation across changing clinical state. The clinical value in such a feature is that it can be a potential indicator of a worsening clinical state. Therefore, recordings of patients’ speech can be a critical component in the monitoring of patients with serious mental illness as these would allow clinicians to more accurately track dynamic signals over time in conjunction with other state-related variables in formal analyses (Cohen et al., 2019). Linguistic variables of patient speech have been shown to contain power in predicting variables of interest in numerous clinical settings. For instance, Elvevåg et al. (2007) showed that levels of incoherence in language can be used in differentiating diagnostic groups and detecting severity in schizophrenia. Recent work in the field of artificial intelligence, and more specifically natural language processing (NLP), has shown that various

cognitive variables can be predicted from language alone (for a review, see Voleti et al., 2019).

Clinicians must integrate information from various behavioral, self-report, and historical sources during the assessment process. The language component of such assessments is just one of a multitude of modalities evaluated. Numerous aspects of serious mental illness can be conveyed in subtleties in a patient’s vocal self-presentation including emotions as expressed in both the sound of voice and the types of words used (Holmlund et al., 2020b), and clinicians may have difficulty noticing and remembering these distinctions. Mental disorders require longitudinal monitoring over several years which is extremely challenging cognitively for clinicians given the different clinical baselines of patients with serious mental illness. Furthermore, it is a difficult task for a human to evaluate the degree to which single modalities of behavior contribute to an individual’s overall mental and cognitive health. Thus, there is a pressing need for automated analytic methods to track and assess mental state in a scalable manner.

As part of a larger study, data were collected through a mobile phone application, the *delta* Mental State Examination (*dMSE*), which administered assessments and collected speech, touch, and self-report data from users in order to track changes in mental state over time (Holmlund et al., 2019; Cohen et al., 2019). Of the 12 tasks administered to the users of the application, we chose a story recall task to answer the question of whether it is possible to predict self-reported measures of positive and negative affect using both speech acoustics and language features.

## 2. Related Work

It has been shown that emotional state can be automatically measured through a person’s speech, both in and out of the laboratory. For instance, the studies of Grimm et al. (2007), Asgari et al. (2014), and more recently, Zhaocheng and

Epps (2018), showed that spontaneous emotion could be accurately predicted from speech. However, all studies took place in a controlled laboratory setting. Our recent work has shown that such analyses are also viable in less controlled settings when tasks are administered remotely via a mobile application (Cheng et al., 2018). In the study of Cheng et al. (2018), 10 speech-based tasks were used to predict self-reported positive and negative affect using only acoustic features. In the current study, we extended this to two modalities: speech acoustics and language, while focusing our analyses only on one single task that does not explicitly elicit emotion (story recall).

One way to measure emotional state is by viewing it as a classification task. In such a task, the goal is to predict speech as belonging to one of the basic categories of emotion (e.g., happiness, anger, fear, etc.). This approach can be problematic since it is hard to get reliable categorization of emotion across evaluators (Mower et al., 2009). While some studies of emotional prediction have attempted to mitigate and overcome this drawback (Steidl et al., 2005), most focus on the prediction of the extent to which certain categories of emotion are present via a continuous representation (Cowie et al., 2012). Thus, the present study employed a regression model to predict emotional state.

Similar to the aforementioned studies, much of the analyses and modeling of behavioral and psychiatric data to date has been in a unimodal manner. For instance, the Interspeech 2018 Computational Paralinguistics Challenge aimed to increase the sensitivity to the non-language information that is conveyed in acoustic properties of speech. Specifically, the self-assessed affect sub-challenge sought to predict the valence of emotions, with the objective of supporting applications for individuals with affective disorders, and for monitoring interactions between therapists and their patients (Schuller et al., 2018).

By focusing solely on acoustic properties, these studies miss the signal contained in natural language features. While the manner in which language is produced at an acoustic level is decidedly important, classic language features have been shown to serve as a window into a person's mental state. For example, natural language features have been shown to accurately predict performance on a story recall task often given as part of the clinical workup in psychiatric settings (Chandler et al., 2019a; Holmlund et al., 2020a). Natural language features have been studied in a range of clinical applications from detecting language impairments in autism to flagging depression in twitter feeds (Goodkind et al., 2018; Coppersmith et al., 2015).

In each study, patient data was reduced to a set of variables to relate to clinical measures of interest. Whether the modality of choice is acoustics, language, reaction time, precision, etc., it has been shown that psychiatric variables of interest can be accurately predicted from unimodal data.

### 3. Data Collection

The *dMSE* mobile phone application was created for the acquisition of cognitive and mental health data of various modalities from both a clinical and non-clinical population. Participants remotely completed sessions, each consisting of a series of 12 tasks, over the course of 3 to 6 days. Such tasks were created to be similar in form and structure to those employed by clinicians in standard

neuropsychological evaluations. One part of each session prompted participants to answer several questions on their emotional well-being by moving a slider to indicate their current level of positive and negative affect (Cohen et al., 2019; Cowan et al., 2019; Le et al., 2018, 2019; more detail in the next section). For the purpose of this study, slider and story recall results were captured by a smart device running the *dMSE* application.

The story recall task prompted participants to listen to a short story and then retell it immediately in as much detail as possible. Stories were all presented verbally and contained two characters, a setting, an action that caused a problem, and a resolution. The content of the stories were designed to be generally well known topics that were emotionally neutral. On average each story was 72 words (SD = 4.6) and each retell was 61.3 words (SD = 21.2 words) and 41.7 words (SD = 21.0 words) for non-clinical participants and clinical participants, respectively. An example story was as follows:

“On Monday morning, the woman woke up more tired than usual. When she walked downstairs to make herself a cup of coffee, she found her husband in the kitchen. She was surprised because he usually left an hour before she woke up. Her husband greeted her and reminded her that daylight savings time was over. Realizing the clocks were wrong, she happily ran upstairs and jumped back into bed.”

The non-clinical subset of our data was composed of 430 sessions that produced valid data from 79 (presumed healthy) undergraduates enrolled in psychology courses at Louisiana State University, yielding 5.4 sessions per student. The clinical subset of our data was composed of 137 sessions that produced valid data from 21 stable clinical participants with a range of serious mental illnesses (schizophrenia, major depressive disorder and bipolar disorder; for details on the assessment procedure in a slightly extended sample, see Holmlund et al., 2020a), yielding 6.5 sessions per participant. This study was approved by the LSU Institutional Review Board (#3618) and participants provided their informed written consent before participation.

### 4. Self-Reported Affect

Each session with the *dMSE* application included sliders to assess general affective states. Participants were prompted as to their emotion on various questions (described below) and asked to indicate their responses on a scale of 0-100. The questions were based on the Positive and Negative Affect Schedule (Watson et al., 1988), which is a tool that measures Positive Affect (PA) and Negative Affect (NA). PA is defined as a state of high enthusiasm, activity, and alertness and NA is defined as a state of distress and unpleasurable engagement. Both are used to quantify mood and are known to be relatively independent of one another. The *dMSE* application contains 7 PA sliders that ask the user to report on personal levels of hopefulness, calmness, appreciation, strength, ability to concentrate, happiness, and levels of energy. Similarly, 8 NA sliders ask the user to self-report on personal levels of anxiety, frustration, fear, sadness, stress, anger, pain, and helplessness. The final self-reported PA and NA values per session is the average of the PA and NA slider responses. The PA results ranged

from 0 to 100 with an average of 74.9 (SD = 21.0) for the clinical group and ranged from 10 to 100 with an average of 64.0 (SD = 17.3) for the non-clinical group. The NA results ranged from 0 to 100 with an average of 29.5 (SD = 23.2) for the clinical group and ranged from 0 to 74 with an average of 26.1 (SD = 17.0) for the non-clinical group.

## 5. Experimental Results

In this study, we generated predictions of positive affect (PA) and negative affect (NA) in a clinical group and a non-clinical group. The predictions were first based on speech features and standard NLP features individually, and then on a combination of these two to answer the question of how well multimodal predictions outperform unimodal predictions.

In each experiment variation, a Support Vector Regression (SVR) model was trained on data from all participants in a group but one, and tested on the set of sessions of each ‘held out’ participant. The SVR model was chosen as it is well-suited for predicting with many continuous independent variables. The SVR parameters were consistent with prior work: a radial basis function (RBF) kernel, degree = 3, cost = 10, eps = 0.2 (Cheng et al., 2018). The reported results are the average correlation between self-reported PA and NA and the predicted PA and NA over all tested participants. Each model was trained with these same parameters as we were more interested in relative improvements in the overall prediction of PA and NA when new features and modalities were introduced than finding the best overall models.

### 5.1 Speech-based results

The first experiment was a re-analysis of prior work (see Cheng et al., 2018 for details). Speech features from the openSMILE audio feature extractor (Eyben et al., 2013) were generated from each story recall response. The openSMILE audio feature extractor is a state of the art package that generates low-level features such as energy, loudness, and voice quality as well as processed statistics of such features such as means, extremes, regressions, and percentiles. We used the entire 2013 ComParE feature set which comprised 6,373 distinct speech features per response (Schuller et al., 2013).

Prior work reported results on the same data and model, but used a 10-fold cross-validation training technique where model parameters were learned using 9 of 10 subsets of the data and tested on the 10th subset for evaluating performance. In contrast, we performed a leave-one-out cross-validation technique where the model parameters were learned using data from all but one participant and tested on the set of sessions from the single ‘held out’ participant. The benefit of this form of cross-validation is that the resulting models more closely resembled how well a fully trained model would perform when applied to new data.

The subsequent analyses were performed on various subsets of the data. All openSMILE features were used in the SVR to predict both PA and NA in the clinical participants as well as in the non-clinical participants. For this part of the analysis, the groups were kept separate. The results of the different variations of the analysis are shown in Table 1. Consistent with prior work (Cheng et al., 2018), we found higher correlations to self-reported affect in the clinical population than the non-clinical population.

	Clinical	Non-clinical
PA	0.40	0.33
NA	0.39	0.36

Table 1: Correlations between self-reported PA/NA and SVR predictions using all openSMILE features.

Many of the features in the openSMILE feature set are highly co-linear, and so receive essentially 0 weighting within the prediction model. While many contribute to the prediction models nearly equally based on small differences in the data, each iteration of the prediction model had a distinct best feature that correlated significantly higher than the rest. For instance, spectral flux and Mel-Frequency Cepstral Coefficients (MFCC) best predicted PA and NA respectively in the clinical group. Similarly, spectral roll-off and MFCC best predicted PA and NA respectively in the non-clinical group.

In addition to models trained and tested on clinical and non-clinical participants separately, a SVR model was trained on data from all clinical participants and tested on data from all non-clinical participants, as well as *vice versa*. The results of this portion of the analysis are detailed in Table 2. When combining data from the two groups of participants, the correlations with PA and NA both significantly decreased. This suggests that the weights of the various speech features used to predict PA and NA have different distributions in the two subsets of participants due to the difference in ranges of self-reported affect.

Training set	Test set	PA	NA
Clinical	Non-clinical	0.16	0.11
Non-clinical	Clinical	0.06	0.08

Table 2: Correlations between self-reported PA/NA and SVR predictions using all openSMILE features when trained on the non-clinical population and tested on clinical, as well as *vice versa*.

Finally, one model with all clinical and all non-clinical data combined was trained and tested in the same leave-one-out manner. The average correlation between the SVR predictions and self-reports of PA was 0.13 and of NA was 0.06. The lower correlations that result from both models that mix clinical and non-clinical data imply that these two populations must be considered independently as their self-reports follow distributions that are distinct from one another.

### 5.2 Language-based results

Since the aim of this study was to test whether the addition of data from a separate modality would improve the ability to predict emotion, we next repeated the speech-based experiments on language-based features to compute a baseline of the power of language features.

Traditionally, story recall is rated manually by assigning points for key words or thematic units correctly recalled. This process can be automated by extracting various task-specific NLP features (e.g., common tokens between the original story and the recall or the cosine distance between the vector representations of the original story and the recall) from each recall response to measure the similarity

between the two (see Chandler et al., 2019 and Holmlund et al., in press for more details).

The audio of each story recall was transcribed by trained humans. Non-task-specific NLP features were computed and modeled against the self-reported affect variables (so as to focus the analyses only on general language features rather than those features that would indicate successful task completion). The NLP feature set included token count, type (unique words) count, type token ratio, content density, mean coherence, standard deviation of coherence, and counts of particular parts of speech such as verbs, nouns, pronouns. Type token ratio is defined as the ratio of word types to word tokens. Content density is defined as the number of verbs, nouns, adjectives, and adverbs to total tokens, or put simply, the ratio of content words to total words. Coherence is computed by comparing adjacent windows in the text for similarity. For the purpose of this study, the window size was chosen to be  $n = 4$  and the similarity metric used was the cosine distance between vector embeddings of the words in each window. The average and standard deviation of similarities of all adjacent windows in a recall were computed. Table 3 shows the results of the different variations of analyses conducted on language-features (which are identical to the variations in the speech-based experiments).

	Clinical	Non-clinical
PA	0.16	0.07
NA	0.28	0.14

Table 3: Correlations between self-reported PA/NA and SVR predictions using all standard NLP features.

Perhaps unsurprisingly, the NLP features were less useful than the speech features in predicting affect. Had the task analyzed been one that specifically seeks to elicit affect in the words spoken, such as the prompt “how are you feeling today?”, we predict that NLP features would be more likely to have a stronger impact on modeling affect.

Finally, a semantic analysis was performed using high-dimensional vector space embeddings of the text. The purpose of the semantic analysis was to measure subtle aspects of language meaning that could be correlated with self-reported affect across different participants. These embeddings operate under the assumption that words that tend to show up in similar contexts are semantically related and thus should be close to each other in a derived vector space. Examples of embedding techniques are simple count based vectorizers (with and without term frequency-inverse document frequency weighting), Latent Semantic Analysis (Landauer and Dumais, 1997), word2vec (Mikolov et al., 2013), and ELMo (Peters et al., 2018). In this experiment, the term frequency-inverse document frequency weighted vectors were the most predictive out of those tested. The term frequency-inverse document frequency weighting accounts for how important a word is to a document based on counts of the word in the entire corpus. Although word2vec and ELMo embeddings are typically regarded as containing more signal in terms of word meaning, they proved unable to predict affect. This is likely due to the fact that the vocabulary of the recall task is limited and thus the increased power of the semantic/syntactic modeling found in these embeddings do not contribute greatly to predicting

affective state. The two variations of our two semantic language-based experiments are detailed below.

First, a k-nearest neighbors (KNN) measure was developed to predict PA and NA based on the affect ratings of the closest recalls in the embedding space to a given recall. Once each recall is projected into an embedding space, the  $k = 6$  (chosen based on the overall best performance on the held out data) closest embeddings of other participant recalls were retrieved and the affect for the session in question is predicted to be some function of those 6. The KNN measure provides an indexing of a participant's PA and NA mental state against the mental state of other participants. For example, if the language used in a response is highly similar to other participants, we can predict that their PA and NA scores would be similar. Second, the recall embeddings were used as input to a SVR model. The same experimental settings were used as in the SVR for speech-based features and the standard NLP-based features. Results of the KNN model and the SVR model are detailed in Table 4. Again, the speech-based features consistently outperform the language-based features.

	Clinical	Non-clinical	Clinical	Non-clinical
	KNN		SVR	
PA	0.11	0.07	0.23	0.18
NA	0.14	0.16	0.31	0.13

Table 4: Correlations between self-reported PA/NA and both KNN predictions and SVR predictions using the recall embedding as input.

### 5.3 Combined results

Finally, to test our hypothesis that the inclusion of multiple modalities in the modeling of self-reported affect is superior to unimodal modeling, we combined the speech-based features with the language-based features and ran the same SVR experiment variations as above.

Combining two modalities improves predictions of self-reported affect by 10-23%. Even though the recall task is a task that is not designed to specifically elicit emotion, the manner in which participants spoke in terms of acoustics and language still contained critical signals indicative of their positive and negative affect.

All features (openSMILE, standard NLP, and recall embedding neighbors) were used in a single SVR model to predict positive and negative affect. Results of this combined model are detailed in Table 5.

	Clinical	Non-clinical
PA	0.44 (10%)	0.38 (15%)
NA	0.48 (23%)	0.40 (11%)

Table 5: Correlations between self-reported PA/NA and SVR predictions using all features, with their relative improvements over unimodal predictions.

## 6. Discussion and Conclusion

In this research, self-reported measures of affect (e.g., PA/NA) were taken separately from the story recall task that was used to predict emotion. Indeed, the nature of a story recall task on neutral stories is not directly designed to elicit emotional state. However, subtle aspects of

emotion were still evident in the language. These results indicate that the approach can derive a fairly stable measure of affect through self-reports that can be predicted in separate tasks. Furthermore, the results indicated that some people's affect levels are easier to predict than others and some types of affect may be easier to predict. For example, Cheng et al. (2018) showed that negative affect was easier to predict than positive affect.

Overall, we have shown that the use of multiple modalities of data in prediction models can lead to a significant increase in power over analyses of a single modality. Speech and language features each contribute independent components that help predict affective state. The speech features contributed more strongly to the predictions which could partially be due to the nature of the tasks used.

Traditionally, unimodal data analyses have been conducted on clinically valuable data as the combination of modalities (and thus data types) can be statistically complex. However, the field of clinical medicine and behavioral science is beginning to see a push for more multimodal analyses. Although the collection of multimodal data is standard in many fields (e.g. neuroimaging; Sun et al., 2020), it is just recently becoming common for multiple modalities to be considered in a single computational model.

Overall, the results show a path towards automatic analysis of patient mental state using both audio and linguistic features. This research has shown that affective state can be predicted from a single task with two modes of communication. In automated assessment of psychiatric variables, it is important to consider multiple modalities of behavior, whether that is within language (considering both acoustic and linguistic data), or beyond, using patient actions, response speed, and other similar variables.

The *d*MSE collected data from a variety of other tasks, including picture descriptions, verbal fluency, memory, tapping, and Stroop tasks. Thus, future research will examine the data from all tasks and run equivalent SVR experiments on features extracted from all sessions of each participant. This opens the possibility of analyzing a combination of speech, language, memory accuracy and touch-based speededness tasks, and could give a more detailed and accurate view of the patients' state, more analogous to what a clinician considers when making decisions. However, with the increase in model complexity, we must be careful, especially in the field of medicine, to not lose the notion of transparency and explainability (Chandler et al., 2019b). Thus, while adding additional modalities and features, it is critical to remember that the goal is not just to build an accurate model, but to understand how that model can be used to inform sound clinical decision-making.

## 7. Acknowledgements

This project was funded by grant 231395 (2014 – 2017) from the Research Council of Norway awarded to Brita Elvevåg. We thank Jared Bernstein and Elizabeth Rosenfeld for their valuable contributions to the project.

## 8. Bibliographical References

Asgari, M., Kiss, G., van Santen, J., Shafraan, I., and Song, X. (2014). Automatic measurement of affective valence and arousal in speech. In *Proceedings of ICASSP*, pp. 965-969.

Chandler, C., Foltz, P.W., Cheng, J., Bernstein, J.C., Rosenfeld, E.P., Cohen, A.S., Holmlund, T.B. and Elvevåg, B. (2019a). Overcoming the bottleneck in traditional assessments of verbal memory: Modeling human ratings and classifying clinical group membership. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pp. 137–147. <http://dx.doi.org/10.18653/v1/W19-3016>

Chandler, C., Foltz, P.W., Elvevåg, B. (2019b). Using Machine Learning in Psychiatry: The Need to Establish a Framework That Nurtures Trustworthiness, *Schizophrenia Bulletin*, Volume 46, Issue 1, January 2020, pp. 11–14, <https://doi.org/10.1093/schbul/sbz105>

Cheng, J., Bernstein, J., Rosenfeld, E., Foltz, P.W., Cohen, A., Holmlund, T., Elvevåg, B. (2018). Modeling Self-Reported and Observed Affect from Speech. 3653-3657. 10.21437/Interspeech.2018-2222.

Cohen, A.S., Fedechko, T.L., Schwartz, E.K., Le, T.P., Foltz, P.W., Bernstein, J., Cheng, J., Holmlund, T.B., and Elvevåg, B. (2019). Ambulatory vocal acoustics, temporal dynamics, and serious mental illness. *Journal of Abnormal Psychology*, 128(2), pp. 97–105. <https://doi.org/10.1037/abn0000397>.

Cohen, A.S., Mitchell, K.R., Docherty, N.M., and Horan W.P. (2016a). Vocal expression in schizophrenia: Less than meets the ear. *Journal of Abnormal Psychology*; 125(2):299-309. doi: 10.1037/abn0000136.

Cohen, A.S., Renshaw, T.L., Mitchell, K.R., and Kim, Y. (2016b). A psychometric investigation of "macroscopic" speech measures for clinical and psychological science. *Behavior Research Methods*. 48(2):475-86. doi: 10.3758/s13428-015-0584-1.

Coppersmith, G., Dredze, M., Harman, C., and Hollingshead, K. (2015). From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. In *Proceedings of the 2nd ACL Workshop on Computational Linguistics and Clinical Psychology*.

Cowan, T., Le, T.P., Elvevåg, B., Foltz, P.W., Tucker, R.P., Holmlund, T.B., Cohen, A.S. (2019). Comparing static and dynamic predictors of risk for hostility in serious mental illness: Preliminary findings. *Schizophrenia Research*. 204, 432-433. doi: 10.1016/j.schres.2018.08.030

Cowie, R., McKeown, G., and Douglas-Cowie, E. (2012). Tracing emotion: an overview. *Int. J. Synth. Emot.* 3, 1–17. doi: 10.4018/jse.2012010101

Elvevåg, B., Foltz, P.W., Weinberger, D.R. and Goldberg, T.E. (2007). Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. *Schizophrenia Research* 93(1-3): pp. 304-316.

Eyben, F., Weninger, F. Gross, F., and Schuller, B. (2013). Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM International Conference on Multimedia*, pp. 835–838.

Goodkind, A., Lee, M., Martin, G.E., Losh, M., and Bicknell, K. (2018). Detecting language impairments in autism: A computational analysis of semi-structured conversations with vector semantics. In *Proceedings of the Society for Computation in Linguistics*, volume 1.

Grimm, M., Kroschel, K., and Narayanan, S.S. (2007). Support vector regression for automatic recognition of

- spontaneous emotions in speech. In *Proceedings of ICASSP*, pp. 1085-1088.
- Holmlund, T.B., Chandler, C., Foltz, P.W., Cohen, A.S., D., Cheng, J., Bernstein, J., Rosenfeld, E., and Elvevåg, B. (in press / 2020a). Applying speech technologies to assess verbal memory in patients with serious mental illness. *npj Digital Medicine*.
- Holmlund, T.B., Fedechko, T.L., Elvevåg, B. & Cohen, A.S. (2020b). Chapter 28: Tracking language in real time in psychosis. In: *A Clinical Introduction to Psychosis: Foundations for Clinical and Neuropsychologists*. Eds. J.C. Badcock & G. Paulik-White. Elsevier.
- Holmlund, T.B., Foltz, P.W., Cohen, A.S., Johansen, H.D., Sigurdson, R., Fugelli, P., Bergsager, D., Cheng, J., Bernstein, J., Rosenfeld, E. & Elvevåg, B. (2019). Moving psychological assessment out of the controlled laboratory setting and into the hands of the individual: Practical challenges. *Psychological Assessment*. 31(3), 292-303. doi: 10.1037/pas0000647
- Landauer, T.K., Dumais, S.T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, vol. 104, no. 2, pp. 211–240.
- Le, T.P., Cowan, T., Schwartz, E.K., Elvevåg, B., Holmlund, T.B., Foltz, P.W., Barkus, E. & Cohen, A.S. (2019). The importance of loneliness in psychotic-like symptoms: Data from three studies. *Psychiatry Research*. doi: 10.1016/j.psychres.2019.112625.
- Le, T.P., Elvevåg, B., Foltz, P.W., Holmlund, T.B., Schwartz, E.K., Cowan, T., Cohen, A.S. (2018). Aggressive urges in schizotypy: Preliminary data from an ambulatory study. *Schizophrenia Research*, 201, 424-425. doi: 10.1016/j.schres.2018.05.045
- Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.
- Mower, E., Metallinou, A., Lee, C., Kazemzadeh, A., Busso, C., Lee, S., et al. (2009). "Interpreting ambiguous emotional expressions," in *The 3rd International Conference on Affective Computing and Intelligent Interaction (Amsterdam)*, 1–8.
- Pennebaker, J.W., Boyd, R.L., Jordan, K. and Blackburn, K. (2015). *The development and psychometric properties of LIWC*. Austin, TX: University of Texas at Austin.
- Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L. (2018). Deep contextualized word representations. *NAACL-HLT*.
- Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., Mortillaro, M., Salamin, H., Polychroniou, A., Valente, F., and Kim, S. (2013). The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Interspeech*, pp. 148–152.
- Schuller, B.W., Steidl, S., Batliner, A., Marschik, P.B., Baumeister, H., Dong, F., Hantke, S., Pokorny, F., Rathner, E.M., Bartl-Pokorny, K.D., Einspieler, C., Zhang, D., Baird, A., Amiriparian, S., Qian, K., Ren, Z., Schmitt, M., Tzirakis, P., and Zafeiriou, S. (2018). "The INTERSPEECH 2018 Computational Paralinguistics Challenge: Atypical & Self-Assessed Affect, Crying & Heart Beats," in *INTERSPEECH*, pp. 1–5.
- Steidl, S., Levit, M., Batliner, A., Noth, E., and Niemann, H. (2005). "Of all things the measure is Man' automatic classification of emotions and inter-labeler consistency," in *ICASSP (Philadelphia, PA)*, pp 317–320.
- Sun, Y., Ayaz, H., and Akansu, A.N. (2020). Multimodal Affective State Assessment Using fNIRS + EEG and Spontaneous Facial Expression. *Brain Sci*. 2020;10(2):E85. doi:10.3390/brainsci10020085
- Voleti, R., Liss, J.M., and Berisha, V. (2020). A Review of Language and Speech Features for Cognitive-Linguistic Assessment. in *IEEE Journal of Selected Topics in Signal Processing - Special Issue on Automatic Assessment of Health Disorders Based on Voice, Speech and Language Processing*.
- Watson, D., Clark, L.A., and Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology*, vol. 54, pp. 1063–1070.
- Zhaocheng, H. and Epps, J. (2018). Prediction of Emotion Change from Speech. *Frontiers in ICT*, vol. 5, pp. 11. doi: 10.3389/fict.2018.00011

# The *RiMotivAzione* dialogue corpus

## Analysing Medical Discourse to Model a Digital Physiotherapist

**Francesca Alloatti, Andrea Bolioli, Alessio Bosca, Mariafrancesca Guadalupi**

CELI - Language Technology, Università degli Studi di Torino  
Turin

{francesca.alloatti, andrea.bolioli, alessio.bosca, mariafrancesca.guadalupi}@celi.it

### Abstract

The *RiMotivAzione* project aims at providing a digital assistant that guides patients in their physiotherapy sessions at home. To properly develop this assistant, we gathered a corpus of dialogues between patients and physiotherapists. In this paper we present a deep and extended analysis of this corpus over different levels of granularity. The linguistic features extracted from the medical discourse were employed to model the *RiMotivAzione* chatbot, which will be experimented with patients at San Camillo Hospital in Venice (Italy).

**Keywords:** dialogue, physiotherapy, chatbot

### 1. Introduction

In the recent years there has been a steep increase in the application of ICT technologies to the healthcare domain (Kafle and Huenerfauth, 2018; Mieskes and Stiegelmayr, 2018; Liao et al., 2019). Specifically, one of these technologies is chatbots, or conversational agents. Most of them are created to help users to better communicate with the clinicians, as well as to help the medical personnel to monitor their patients (Laranjo da Silva et al., 2018). These chatbots are developed with different frameworks involving various techniques (reinforcement learning, pattern matching, etc.) and they are deployed throughout a range of platforms (Montenegro et al., 2019). Despite the abundance of systems, little to no description is provided about the language employed by the chatbots. By "language" here we mean the combination of words, phrases, tone and pragmatic features employed while giving instructions or providing any kind of medical assistance. A lot of attention has been paid to the linguistic features clinicians apply when talking to patients (Ferguson, 2012), therefore it is logical that the same focus should be applied when chatbots are the ones conversing with the patients. A correct use of medical language has been proved to be essential to a positive outcome of the treatment path (Hull, 2016) and a conversational agent should use the same terminology used by doctors and nurses.

In the *RiMotivAzione*<sup>1</sup> research project, a conversational interface is integrated with a visual App and a wearable device equipped with motor sensors. The project aims at assisting elders who suffered from a stroke and are under treatment for upper limb motor rehabilitation (Bolioli et al., 2019). The chatbot works as a virtual physical therapist guiding the patients through the exercises, giving advice and asking for information about the person's well-being. Given the aforementioned importance of the use of correct language in the medical domain, the interface has been modeled after therapists' real linguistic behavior: a corpus of conversation between doctors and patients was

collected, transcribed and studied to retrieve information about linguistic communication in the physical rehabilitation domain. Preliminary results can be found in Bolioli et al. (2019), while in this work we present the data in more detail. We conduct a deeper analysis at various levels of granularity and provide more insight into the features of the physiotherapist-patient communication.

To the best of our knowledge, this is the sole corpus that deals with linguistic features employed in a specific medical setting - physical therapy sessions - in Italian. The corpus is not publicly released due to privacy reasons. It can be obtained for research purposes by writing to the authors.

### 2. Related Work

An analysis of related work unveiled various studies that share similarities with this one. The most similar one gathered a corpus of conversations between therapists and patients and analyzed it (Chang et al., 2013), although the language used is Korean. The analysis highlights some interesting features that can also be found in our corpus, such as the imbalance between the number of patients' utterances and the doctors' ones. They have also taken into consideration non-verbal behavior to measure empathy. Their goal was in fact to improve empathetic communication, while ours is to model a chatbot. Chaoua et al. (2018) also concerns analysis of patients-therapists conversations, although their setting is a psychological one, and their goal is topic detection and extraction. In a similar way, Jin (2018) focus solely on the analysis of small talk. Mieskes and Stiegelmayr (2018) inspect data from psychotherapy sessions with the aim of identifying what constitutes a sign of cooperation between the two participants. For this reason, their analysis is mainly qualitative. Finally, Wang et al. (2018) gather data about conversations in the pediatric domain.

Other works have different goals, such as producing a different metric to evaluate ASR system (Kafle and Huenerfauth, 2018), or even to model a dialog system (Gilmartin et al., 2018), although this last study does not focus on conversational interfaces in the healthcare domain.

<sup>1</sup>*RiMotivAzione* is a two-year Research and Innovation project supported by POR FESR 2014-2020 Regione Piemonte

Regarding the annotation and analysis of the corpus, we consulted the work by Shelley Staples (2016), in which certain linguistic features are extracted from corpora of medical discourse given their importance and ability to represent the quality of the exchange between doctors and patients. We analyzed our corpus identifying the features that we deemed relevant for the specific domain of physical rehabilitation. On the other hand, the annotation of speech acts poses a different kind of challenge. It has been tackled by various means and more recently through the employ of automatic systems (Basile and Novielli, 2018). However, no specific work focus on the automatic annotation of acts for the medical discourse, which may require a different set of tags and approach.

### 3. The *RiMotivAzione* Corpus

The *RiMotivAzione* corpus contains dialogues between a physician and a patient during the course of physiotherapy sessions. The people involved in the recording are two patients and three physiotherapists. Both patients are elderly (more than 60 years old) and males. Only one of the clinicians took care of both patients, while the other two were assigned to just one care recipient. The patients were selected by the research team at IRCSS San Camillo Hospital in Venice based on some preliminary tests. These tests aimed at identifying patients that could take part in the study by having certain characteristics: for instance, their speech needed to be sufficiently clear and they needed to be proficient in the Italian language. This was meant to exclude people who speak mainly in their own regional dialect, which is usually not intelligible from people from other parts of Italy.

Both patients signed an informed consent to be recorded and to have their data handled according to the current privacy laws.

#### 3.1. Setting and Corpus Features

The sessions were recorded at IRCSS San Camillo Hospital. They were taped by means of a camera and the audio was extracted from the videos. This is due to some ambient conditions of the room where the sessions took place: high ceiling and the presence of temporary drywall generated a lot of noise and rumbling sounds from the rooms nearby. Professional recorders - that we employed at first - are very sensitive and captured each vibration, generating too much disturbance for the automatic transcription engine. Moreover, we did not want to use wearable microphones, since they would have disrupted the physiotherapy conditions.

The audio extracted from the video, in mono compression, has a lower quality but also presents less background noise. No additional information useful to the study could be deducted from the videos and they posed a privacy problem, therefore the visual track was not included in the corpus. The original files were transferred from the Hospital to the rest of the research team by mean of a private repository, in order to safely exchange data regarding the patients.

Each session is composed of three different stages, while only the last session contains four. The clinicians stopped the recording at each change of stage, so that the transition from one another was easier to understand even after the

event. The stages are *Reception*, *Calibration* and *Therapy*. An addition *Screening* stage can be found mainly during the first or last session. The mean duration of each session is one hour.

During the transcription part each file pertaining to a different stage was joined in a chronological order, creating two main collections, one for each patient. The data available for each collection can be found in Table 1.

Data	Format
Unique ID per line	Integer
Transcription	String
Annotation	String
Timestamp	Date Time
Session stage	String
Session number	Integer
Name of the clinician	String
Patient anonymous ID	String
Age of the patient	Integer
Sex of the patient	String

Table 1: Corpus data and its format.

Additional information is available for each collection, such as the time span of the recordings, the total number of sessions and the total number of lines, whereas "lines" is used as a synonym for "turns". For the first patient collection:

- Time span: December the 3rd, 2018 to December the 20th, 2018
- Number of sessions: 14
- Number of lines: 3373

For the second patient collection:

- Time span: February the 25th, 2019 to April the 8th, 2019
- Number of sessions: 16
- Number of lines: 4293

#### 3.2. Transcription Methodology

The corpus was produced by means of a semi automatic approach; we manually revised the textual output created by an automatic transcription engine in order to correct the problems emerged during the transcription and to obtain a dialogue corpus with a high degree of accuracy.

The automatic transcription was carried out with a transcription engine developed for commercial purposes. To adapt it to our need, the engine was fine tuned to a portion of the corpus data. This pre-processing was essential to improve the final performance of the system. However, the outcome still presented a significant Word Error Rate over the entire corpus, such that ex post processing (e.g. post transcription fine-tuning) did not produce meaningful results.

The rationale of this poor performance is to be found in the intrinsic nature of such data - dialogues in a real setting

- which is inherently more difficult than standard corpora. To this matter, we tested various ASR engines and obtained similar results. Moreover, the patients spoke Italian with a heavy accent, and even though they were asked not to use dialect, sometimes they slipped some words in their spoken flow without realizing it. None of these difficulties could be addressed automatically, so the entire corpus was manually revised.

### 3.3. Manual Revision

The output of the system was manually revised and corrected following Savy's guidelines (2005) for transcription of spoken Italian. We added proper punctuation to help interpret the meaning of the sentences and marked with a specific tag `Unclear` the parts that were either unintelligible or in dialect. Since the `Unclear` tag could be applied to single words or to entire sentences, the Word Error Rate proved to be an unreliable metric: some words could not be understood because of the dialect, or entire sentences were muffled by background noise such that even a human transcriber could not understand them. For the first patient, the `Unclear` tag appears 238 times, while for the second one 145 times. Proper names of patients and doctors were anonymized to preserve privacy.

Overlapping contents between the two speakers and pauses were not specifically marked or tagged, as it was not relevant to our study.

## 4. Corpus Analysis

Each collection properly assembled and corrected was analyzed with the goal of obtaining objective measurements of the physiotherapist linguistic behaviour. The features extracted were to help model the chatbot ability to efficiently communicate with the patients.

Even though the major focus is on the physiotherapist's part of the dialogue, we also analyzed and discussed the patient's speech. The goal was to highlight how he reacted to certain linguistic stimuli given by the doctor, if there were certain words he did not understand, what were his expectations - if he ever expressed any, etc. Since a chatbot is inherently less smart than a human therapist, we needed to predict any possible difficulty conveyed by the patient so to address it properly through an efficient conversational design.

The analysis was carried out on two levels of granularity: in the more detailed one, we considered the single token up to its morphemes, as well as the dependencies in a single sentence. This analysis was conducted with open source *StanfordNLP* library for Python <sup>2</sup>. On a broader level, we annotated each turn pertaining to the patient or the therapist with a dialogue act tag (ore more than one, if necessary). We employed the RIAS tagset, which is specific to the domain of medical discourse and thus allowed for a more precise definition of each dialogue act. RIAS was developed for encoding conversation in the medical domain in 1991 by Debra Roter et al. (1991) (2002) and it has been applied to various settings, e.g. to annotate exchanges between doctors and oncological patients, for psychotherapy sessions

<sup>2</sup><https://stanfordnlp.github.io/stanfordnlp/>

or even when the dialogue takes place between clients and pharmacists (Roter et al., 2017).

### 4.1. The RIAS tagset

The RIAS tagset has been designed to cover all the speech acts that could appear in the medical discourse. It contains 29 categories grouped in four macro-categories called Medical Interview Functions (MIF). These macro-categories are *Data Gathering*, *Information Exchange*, *Emotional Expression and Responsiveness*, *Partnership Building* and *Activation*. Table 2 contains an excerpt of the complete list of categories. For brevity reasons, we present only the ones that occur at least 200 times in the corpus, together with real examples taken from the dialogues. The examples are translated for the purpose of this paper and are selected for their clarity with respect to the category.

All the categories defined in Roter et al. (2017) were employed in the annotation. Nonetheless, not all of the tags always applied completely to the situation, or some tags were under-represented in this corpus compared to other studies: for instance, the tag `Concerns` was assigned to fewer turns, since these patients did not present a critical clinical situation and their chance of recovery was good (in contrast to other situation, such as an oncological one). Two additional tags were included to cover the entirety of speech acts in the dialogues: `Unclear` and `Technical problems`. The first one was used to tag incomplete sentences, the ones where the original audio was too compromised to understand the words, or the patients talked in their dialect. The `Unclear` tag was also employed in cases where the speech referred to the context in real time, making the general meaning impossible to retrieve for the annotator just by listening to the audio. The video track did not provide any help in resolving these matter. On the other hand, the `Technical problems` tag applied where the armband device used by the patients presented some issue. The two speakers may then discuss the subject of technological devices, which went beyond the tags presented in the RIAS tagset.

### 4.2. The Annotation Process

While the detailed analysis was conducted automatically with the help of *StanfordNLP* library, the speech act annotation was carried out manually. Three annotators took part in the work: one of them also served the purpose of super-annotator in case of disagreement. All the annotators have a formal education in Linguistics and they are aware of standards and annotation procedure regarding linguistic corpora. Each dialogue turn may contain more than one sentence and it may express more than one speech act. Therefore, a single turn can be tagged with two or more tags.

Inter-annotator agreement between two of the workers was calculated at ( $k = 0.63$ ) according to Cohen's score. In case of disagreement, which happened in about 25% of the data, the super-annotator worked as a conciliator until all the annotators agreed to a final decision.

RIAS code	Example from the corpus
Social talk	non vedevo l'ora di venirla a trovare. <i>I couldn't wait to come and meet you.</i>
Directions	per scendere chiudo, per salire apro la mano. <i>to lower the hand I close it, to lift it up I open it.</i>
Agreements	esatto, perché lo abbiamo registrato proprio così. <i>exactly, we set it this way.</i>
Medical condition	un po', poco, fastidio piú che male. <i>just a bit, it bothers me rather than hurting me.</i>
Approvals	bravissimo. <i>very good.</i>
Unclear	[dialetto veneto] vara! <i>[venetian dialect] look!</i>
Therapeutic regimen	venerdì faremo la parte clinica ti faró io la scala di valutazione. <i>Friday we will do the clinical part I'll make the evaluation scale for you.</i>
Jokes and laughter	ci vediamo domani, è piú una minaccia che un invito. <i>see you tomorrow, sounds more like a threat than an invitation.</i>
Asking for understanding	vorrei portarla così, hai capito? <i>I'd like to bring it down like this, you see?</i>
Checking for understanding	chiudo le dita, così? <i>do I close my fingers, like this?</i>
Concerns	sei sicura che funziona? <i>are you sure it works?</i>
CeQ Medical condition	a fare gli esercizi non ha dolore? <i>do you feel pain while you do the exercises?</i>

Table 2: Frequent RIAS codes. Each code is presented with an explanatory example taken from the corpus.

## 5. Results

The complete *RiMotivAzione* corpus contains about 98778 tokens. The total number of dialogue turns is 7670: 3377 lines for Patient 1, and 4293 lines for Patient 2. To have a first overview on the exchanges between physiotherapists and patients, we present the number of types, tokens, and ratio between types and tokens (defined as the Lexical Richness Index), as well as the amount of questions for the two parts of the corpus (Table 3 and Table 4). Although the two patients do not present the same behavior regarding the number of questions, it can be noticed that the Lexical Richness Index ranges from 0 to 1 in both cases and it has a lower value for the physiotherapists' discourse. This means that doctors do not deploy a large and differentiated terminology, instead they rather stick to a certain script (which is usually an official one that has been validated by the hospital). On the other hand, patients may chat more about personal subjects since they do not need to comply to official clinical procedures.

From a stricter quantitative perspective, the patient produces less words. If we cross this information with the Lexical Richness Index data, we can infer that the patients may talk less but he can roam more freely from one topic to another. In fact, he may chat about some interest of his or about his personal life. This behavior is not only allowed but also encouraged, because it serves as a conversation enhancer and it produces health benefits for the patient, as mentioned in (Delany et al., 2010; Edwards et al., 2004) and in contrast with other findings in literature (Maynard and Hudak, 2008).

Data	Patient 1	Clinician
Types	2065	3017
Tokens	10533	39305
Lexical Richness Index	0.19	0.07
Questions	40	667

Table 3: Data from Patient 1 sessions.

Data	Patient 2	Clinician
Types	2451	2406
Tokens	18233	30707
Lexical Richness Index	0.13	0.07
Questions	380	805

Table 4: Data from Patient 2 sessions.

### 5.1. Part-of-Speech Analysis

A deeper analysis was conducted with respect to the part-of-speech of each token. Complete results can be found in Figure 1. For each patient, two physiotherapists conducted the sessions, according to their availability. The therapist number two intervened for both patients.

We focus on the most abundant PoS tag: verbs. Verbs are indeed the core of a sentence in a language such as Italian, and they express the essence of the action. In detail, verbs at the plural form were deemed to be particularly significant, in the light of their abundance. Table 5 and Table 6 highlight the usage of such verbs from both patients and clinicians. Although the values for Patient 2 are higher, in

	P1	T1	T2	T1+T2	P2	T2	T3	T2+T3	TOT_PATIENTS	TOT_THERAPISTS
ADJ	3403	2931	2849	5780	2172	2674	2579	5253	5575	11033
ADP	5227	4156	4028	8182	3499	3535	3742	7277	8726	15459
ADV	6556	5313	5290	10603	5122	5720	5357	11077	11678	28236
AUX	2182	1496	1445	2941	2018	1469	1470	2939	4200	5890
CCONJ	1288	1063	1065	2128	1255	1059	1001	2060	2543	4188
DET	6207	4990	4661	9651	3948	3948	4024	7972	10155	17623
INTJ	1711	833	975	1808	1288	1432	1436	2868	2999	4676
NDUN	8478	6942	6522	13464	5847	5751	5953	11704	14325	25168
NUM	496	305	288	593	381	247	356	603	877	1196
PRON	3696	2656	2571	5227	3574	2443	2536	4979	7270	10206
PROPN	268	121	127	248	295	124	167	291	563	539
SCONJ	1008	803	790	1593	1166	811	812	1623	2174	3216
VERB	10419	8431	8696	17127	7363	7250	7023	14273	17782	31400
X	2352	1662	1605	3267	2120	1552	1557	3109	4472	6376

Figure 1: Complete results of the PoS tagging analysis. *P* refers to Patient, while *T* to the Physiotherapist.

both cases the one who largely employs verbs is the doctor. This is coherent with the greater use of nouns by the patient. Verbs are often in the indicative mood, which means that most sentences are main clauses. Main clauses are clearer, easier to process from a neurological point of view and they would serve better in the medical domain, where clarity is of paramount importance (Fengler et al., 2016). To corroborate these considerations about the doctors’ manner of speaking, we cross this data with the analysis conducted on the dialogue acts. Most verbs in the indicative mood from the physiotherapists’ discourse are embedded in turns tagged as *Directions*, where the clinician explains to the patient what to do in order to perform an exercise properly. The use of indicative can be expected while giving directions, since it allows for a clear discourse without any nested subordinate, but at the same time it is more polite than the imperative mood.

Verbs at the plural form	1185
Indicative mood	1019
Patient	182
Physiotherapists	1003
Embedded in <i>Directions</i>	846

Table 5: Verbs in Patient 1 sessions.

Verbs at the plural form	1381
Indicative mood	1292
Patient	492
Physiotherapists	1168
Embedded in <i>Directions</i>	969

Table 6: Verbs in Patient 2 sessions.

The second most frequent PoS tag are nouns. However, to the purpose of this study they did not represent an interesting area of analysis. Nouns may pertain to a broad variety of subjects, even some unrelated to the physiotherapy session. Patient 1, for instance, chatted about a hobby of his (motorcycles), therefore some nouns pertained to that area, which is not useful when analysing a medical discourse. This chatty behavior is quite frequent in the elderly, since they tend to talk about a variety of subjects even if they are not related to the context (Kallirroi et al., 2010).

On the other hand, the adjectives, especially the ones employed by the doctors, proved to be an interesting feature. Table 7 lists the the ten most frequent adjectives used by

the physiotherapists. The frequency is computed over the totality of the corpus.

Rank	Frequency	ADJ
1	376	bravo <i>good</i>
2	366	bravissimo <i>very good</i>
3	271	pronto <i>ready</i>
4	244	esatto <i>exact</i>
5	159	ottimo <i>great</i>
6	126	attento <i>careful</i>
7	120	alto <i>high</i>
8	120	fermo <i>still</i>
9	102	giusto <i>right</i>
10	99	rilassato <i>relaxed</i>

Table 7: Adjectives employed by the physiotherapists.

Most adjectives express a positive sentiment, while the rest concerns technical aspects of the therapy (such as “high”, employed while giving instructions to the patients on how to position the wrist). When modeling the chatbot such considerations are fundamental, because if a patients expect to be praised and encouraged during the sessions through the use of certain words, a digital assistant should behave the same way.

## 5.2. Analysis of the Dialogue Acts

The first analysis carried out on the dialogue acts aimed at identifying the quantity of each tag in the totality of the corpus. The distribution of the tags was plotted on a logarithmic scale for patients and physiotherapist (Figure 2). The utterances tagged as *Social talk* are abundant for both speakers, followed by the *Directions* sentences. Even though we could expect tags related to the physiotherapy sessions to be more copious, only the latter is directly connected to the therapy. *Social talk* is very present because it serves its purpose during the sessions: talking about personal matters, doing small talk has a positive ef-

fect on the medical outcome (Gard and Gyllenstein, 2000). Unfortunately, the tag `Unclear` is also quite abundant, although it must be considered that a single turn may have multiple tags, and the `Unclear` may refer to just a word or a part of that line, not the entire exchange.

Given the goal of the study - to model a chatbot after the doctors' way of talking - we focus on the tags that appear more in the physiotherapists' discourse. The great quantity of `Approval` tags is coherent with the analysis conducted on adjectives from the previous subsection: praises such as "alright, that was great" (Patient 2, line 224) or "very good, now close your hand in a fist" (Patient 2, line 115) contain the aforementioned adjectives and are indeed tagged as `Approval`.

The tag `CEQ - Medical conditions` refers to questions through which the physiotherapist checks on the patient's well-being. There are numerous utterances under this tag, which means that the clinicians often check on the status of their care recipient (e.g. "do you feel comfortable?" in Patient 1, line 1). The same considerations can be made for the `Asking for understanding` tag, where the doctor makes sure the patient is onboard with the therapy. `Concerns` are particularly abundant for the patient, as it is to be expected. No specific trend of tags could be found across the corpus, which means that the `Concerns` are distributed all over the sessions, and they do not increase nor decrease along the therapy path.

## 6. Discussion

The analysis over the physiotherapists' discourse revealed some interesting features. First of all, the great presence of verbs in a plural form. Most of these verbs are used together with the *we* pronoun, which suggests a cooperation between the patient and the doctor. Empathy is a fundamental component during the sessions which allows for a quicker healing process (Palma and Sidoti, 2019). The most frequent adjectives highlighted by the analysis are functional to the same pattern of action. The physiotherapists praise the patient's effort and employ a communication strategy that puts the two of them on the same level, eliminating any hierarchy that may cause discomfort.

Some dialogue acts also comply to this strategy: the abundance of `Social talk` and `Approvals` tags imply that digressing from the strict subject of the therapy serves a purpose in the medical discourse. If a patient is chatting and his efforts are reckoned, he may relax more, feel less pain and therefore find the physiotherapy session less hard. All of these linguistic features represent valuable instruction on how a digital assistant for physiotherapy should be developed. However, not all the information from the dialogue can be mapped in the chatbot. The `Jokes` and `laughter` tag, for instance, refers to the use of irony (particularly heavy for Patient 2) and other jokes made from both speakers. Given the contextual nature of laughing matters, it would be unwise for a digital system to mimic such linguistic behavior. We said before that some of the `Unclear` tags are used when the patient is speaking in dialect. Such a feature, although very interesting from a sociolinguistic point of view, cannot be used when predicting the possible input. Dialect does not get properly transcribed

by the ASR systems and it can be tricky to interpret even afterwards.

The *RiMotivAzione* chatbot should be clear when giving instructions but not stiff. It needs to check quite often on the patients well-being, making sure what is their level of pain and how are they handling it. It has to be able to correct the patient when he is performing the exercise wrong, but it should also praise him when he is getting good results. It has to be able to conduct small talk, but not to make jokes or comments that may result inappropriate out of context.

## 7. Conclusions

We gathered a corpus of dialogues between patients and physiotherapists recorded during real therapy sessions. The aim is to analyse the medical discourse and to extract relevant linguistic features at different levels of granularity. We first considered the single words and explored the value of the most frequent parts-of-speech: verbs and adjectives. Nouns were deemed not to be useful. Then, we annotated the dialogues with the `RIAS` tagset, a group of tags created to annotate medical discourse. The annotation and the subsequent analysis produced interesting results: the most frequent tags do not strictly concern the therapy, they rather serve as a psychological support for the patient. The analysis was expanded and deepened with respect to previous work (Bolioli et al., 2019).

All of these features have been incorporated in the development of the chatbot. The *RiMotivAzione* digital assistant is able to explain the exercises and provide praises when they are executed correctly. It can also check on the patient status and gather feedback about his level of pain. The chatbot, together with the smart wristband and the app will be experimented in San Camillo Hospital. After the experimentation, patients will be able to provide validation over various aspects of the project, including the language employed by the chatbot.

### 7.1. Future Work

In the future we plan on expanding the corpus. Unfortunately, only two patients could be enrolled in the present study, while we would like to add supplementary contributions to make the corpus more robust. More data could also be useful to conduct tasks such as automatic annotation and analysis of the tags distribution.

Future work will also embody the results from the experimentation with the patients in San Camillo Hospital, as well as more details about the interaction between the chatbot, the app and the smart wristband.

## 8. Acknowledgements

*RiMotivAzione* is a two-year Research and Innovation project supported by POR FESR 20142020 Regione Piemonte. The partners of the project are Koiné Sistemi, CELI, IRCCS Fondazione Ospedale San Camillo, Synesthesia, Istituto Italiano di Tecnologia (IIT) and Morecognition. The smart wristband was developed by Morecognition while Synesthesia developed the app that coordinates the wristband with the chatbot. We thank Andrea Turolla and Giorgia Pregolato who enrolled the patients and will conduct the experimentation.

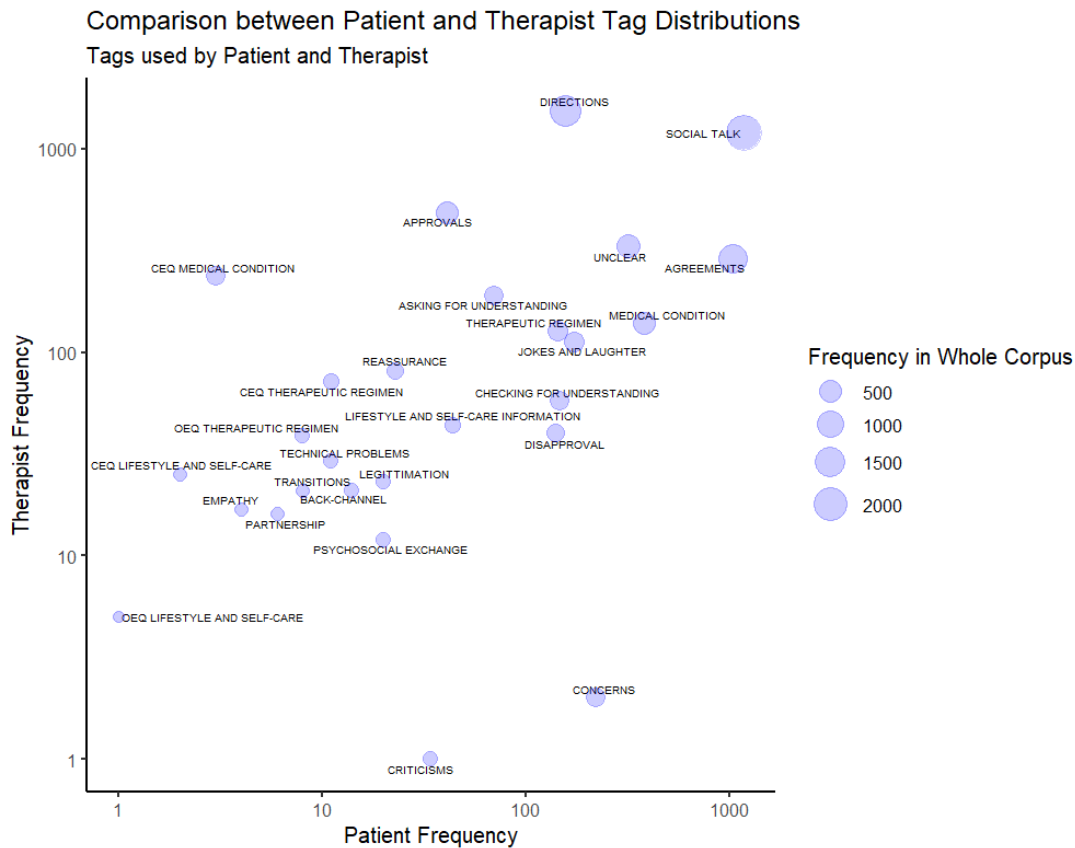


Figure 2: Distribution of dialogue tags across the entire corpus.

## 9. Bibliographical References

- Basile, P. and Novielli, N. . (2018). Overview of the evalita 2018 italian speech act labeling (ilisten) task. In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*.
- Bolioli, A., Alloatti, F., Guadalupi, M., Lanzi, R. I., Pregnolato, G., and Turolla, A. (2019). How do Physiotherapists and Patients talk? Developing and annotating RiMotivAzione dialogue corpus. In *Proceedings of the Sixth Italian Conference on Computational Linguistics*, Bari, Italy, November 12-15, 2019.
- Chang, C., Park, B., and Kim, S. (2013). Conversational analysis of medical discourse in rehabilitation: A study in korea. *The journal of spinal cord medicine*, 36(1):24–30.
- Chaoua, I., Recupero, D. R., Consoli, S., Harma, A., and Helaoui, R. (2018). Detecting and tracking ongoing topics in psychotherapeutic conversations. *AIH@IJCAI*, pages 97–108.
- Delany, C., Edwards, I., Jensen, G., and Skinner, E. (2010). Closing the gap between ethics knowledge and practice through active engagement: an applied model of physical therapy ethics. *Physical Therapy*, 90(7):1068–1078.
- Edwards, I., Jones, M., Carr, J., Braunack-Mayer, A., and Jensen, G. (2004). Clinical reasoning strategies in physical therapy. *Physical Therapy*, 84(4):312–330.
- Fengler, A., Meyer, L., and Freiderici, A. D. (2016). How the brain attunes to sentence processing: Relating behavior, structure, and function. *Neuroimage*, 129:268–278.
- Ferguson, G., (2012). *English for Medical Purposes*, pages 243–261. John Wiley and Sons.
- Gard, G. and Gyllenstein, A. L. (2000). The importance of emotions in physiotherapeutic practice. *Physical Therapy Reviews*, 5(3):155–160.
- Gilmartin, E., Vogel, C., and Campbell, N. (2018). Chats and Chunks: Annotation and Analysis of Multiparty Long Casual Conversations. In Nicoletta Calzolari (Conference chair), et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).
- Hull, M. (2016). Medical language proficiency: A discussion of interprofessional language competencies and potential for patient risk. *International Journal of Nursing Studies*, 54:158–172.
- Jin, Y. (2018). Small talk in medical conversations: Data from china. *Journal of Pragmatics*, 134:31–44.
- Kafle, S. and Huenerfauth, M. (2018). A Corpus for Modeling Word Importance in Spoken Dialogue Transcripts. In Nicoletta Calzolari (Conference chair), et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*,

- Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).
- Kallirroi, G., Wolters, M., Moore, J. D., and Logie, R. H. (2010). The match corpus: A corpus of older and younger users' interactions with spoken dialogue systems. *Language Resources and Evaluation*, 44(3):221–261.
- Laranjo da Silva, L., Dunn, A. G., Tong, H. L., Kocaballi, A. B., and Chen, J. (2018). Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association : JAMIA*, 25:1248–1258.
- Liao, Q. V., Wang, Y.-C., Bickmore, T., Fung, P., Grudin, J., Yu, Z., and Zhou, M. (2019). Human-agent communication: Connecting research and development in hci and ai. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*, pages 122–126, New York, NY, USA. Association for Computing Machinery.
- Maynard, D. and Hudak, P. (2008). Small talk, high stakes: Interactional disattentiveness in the context of prosocial doctor-patient interaction. *Language in Society*, 37:661–668.
- Mieskes, M. and Stieglmayr, A. (2018). Preparing Data from Psychotherapy for Natural Language Processing. In Nicoletta Calzolari (Conference chair), et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).
- Montenegro, J. L. Z., da Costa, C. A., and da Rosa Righi, R. (2019). Survey of conversational agents in health. *Expert Systems with Applications*, 129:56–67.
- Palma, S. and Sidoti, E. (2019). La comunicazione nei processi di cura. *COMUNIT IMPERFET*, 46(4):243–251.
- Roter, D. and Larson, S. (2002). The roter interaction analysis system (rias): utility and flexibility for analysis of medical interactions. *Patient education and counseling*, pages 128–132.
- Roter, D., Isenberg, S., and Czaplicki, L. (2017). The roter interaction analysis system: Applicability within the context of cancer and palliative care. *Oxford Textbook of Communication in Oncology and Palliative Care*, pages 717–726.
- Roter, D. (1991). *The Roter method of interaction process analysis (RIAS manual)*. The Johns Hopkins University, Baltimore.
- Savy, R. (2005). *Specifiche per la trascrizione ortografica annotata dei testi in Italiano Parlato. Analisi di un dialogo*. Liguori, Napoli.
- Staples, S., (2016). *Identifying Linguistic Features of Medical Interactions: A Register Analysis*. Palgrave Macmillan.
- Wang, N., Song, Y., and Xia, F. (2018). Constructing a Chinese Medical Conversation Corpus Annotated with Conversational Structures and Actions. In Nicoletta Calzolari (Conference chair), et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).

# Automatic Quantitative Prediction of Severity in Fluent Aphasia Using Sentence Representation Similarity

**Katherine Ann Dunfield, Günter Neumann**

German Research Center for Artificial Intelligence (DFKI)  
Saarbrücken, Germany  
{katherine.dunfield, guenter.neumann}@dfki.de

## Abstract

Aphasia is a neurological language disorder that can severely impair a person’s language production or comprehension abilities. Due to the nature of impaired comprehension, as well as the lack of substantial annotated data of aphasic speech, quantitative measures of comprehension ability in aphasic individuals are not easily obtained directly from speech. Thus, the severity of some fluent aphasia types has remained difficult to automatically assess. We investigate six proposed features to capture symptoms of fluent aphasia — three of which are focused on aspects of impaired comprehension ability, and evaluate them on their ability to model aphasia severity. To combat the issue of data sparsity, we exploit the dissimilarity between aphasic and healthy speech by leveraging word and sentence representations from a large corpus of non-aphasic speech, with the hypothesis that conversational dialogue contains implicit signifiers of comprehension. We compare results obtained using different regression models, and present proposed feature sets which correlate (best Pearson  $\rho = 0.619$ ) with Western Aphasia Battery-Revised Aphasia Quotient (WAB-R AQ). Our experiments further demonstrate that we can achieve an improvement over a baseline through the addition of the proposed features for both WAB-R AQ prediction and Auditory-Verbal Comprehension WAB sub-test score prediction.

**Keywords:** Aphasia, Quantitative Severity Prediction, Comprehension Impairment, Low Resource

## 1. Introduction

Aphasia is a neurological language disorder, often resulting from stroke, that is characterized by language impairments that affect the production or comprehension of spoken language. Although studies have found that frequent and intensive post-stroke rehabilitation for aphasia is most beneficial in the acute stage following a stroke (Laska et al., 2011; Bhogal et al., 2003), persons with aphasia (PWA) are not always able to obtain the intensity of treatment they need during this stage, or even in later chronic stages. Depending on the location and size of the brain damage acquired, aphasia type and severity can be incredibly variable, where a PWA may exhibit a wide range of language deficits and symptoms. These variations can make it difficult to uniformly extract features of aphasia, particularly symptoms that are not explicitly expressed in a PWA’s speech, such as comprehension impairments. Nonetheless, accurate predictive modelling of aphasia severity offers possibilities in facilitating more personalized and intensive treatment for aphasic patients.

Within the field of natural language processing, considerable previous work has been done in both detecting aphasia and adapting existing technology to be of better used by PWAs (Adams et al., 2017; Le et al., 2017; Fraser et al., 2014a; Fraser et al., 2014b; Thomas et al., 2005; Fraser et al., 2014c). However, due to the differences in nature of aphasia types, the primary focus of this research has been on non-fluent aphasias, which are distinguished predominately by observable production errors, and are therefore easier to obtain from narrative elicitations. Fluent aphasia, on the other hand, especially fluent aphasias noted by impairments in comprehension and semantically incoherent speech, are more difficult to observe outside of a conversational setting where confirmation of whether a lapse

in comprehension has occurred can be established.

Quality data for aphasia speech is rather limited, since it takes comparatively more time and effort to find and record post-stroke aphasic speech than it does for other types of spoken language data. Likewise, because data regarding aphasia deals with real people and often needs to include significant real-life data to be useful, privacy issues become a major concern, as is often the case in medical data. The basis of the approach in extracting features for aphasia without significant training data is to leverage the dissimilarity of aphasic speech with abundant non-aphasic training data, using a few proposed methods. With the assumption that the non-aphasic data offers a survey of healthy speech, deviation from this speech can be viewed as a symptom of aphasia. By using non-aphasic speech as a baseline and computing features through dissimilarity, we create an approach that does not rely on sizeable training data of aphasic speech.

A defining characteristic of many fluent-aphasia types is a lack of understanding of both written and auditory input. As previously mentioned, much work has been performed on identifying features suitable for non-fluent aphasia, focusing on relatively surface level features, such as word frequency and speed of speech. Fluent aphasia, on the other hand, will often differ less from non-aphasic speech than the non-fluent varieties of aphasia, and is instead characterized by a lack of semantic coherency and deteriorated comprehension abilities. Therefore, to capture comprehension impairments in conversational discourse, we assume that comprehension errors often result in inappropriate responses to comments in conversational discourse.

It can be argued that since aphasia severity is expressed in a multitude of ways, achieving reliable modelling of aphasia rehabilitation depends on the availability of data that cov-

ers a sufficient range of aphasia types and symptoms, and a method of better capturing the more implicit symptoms of aphasia. In this work, we propose an investigation into a set of features, specifically selected to capture the primarily distinctive features of fluent aphasia types. These features may be extracted using state-of-the-art methods in natural language processing that allow for analysis of the semantic content of speech. We therefore present three main contributions aimed to overcome issue related to data sparsity and implicit feature extraction: a method of automatically assessing comprehension ability in conversational discourse, leveraging the dissimilarity between healthy and aphasic data to estimate the degree of severity, and utilizing a metric learning approach to capture the likelihood of an aphasic utterance as to track aphasia severity in a measurable way.

## 2. Related Work

Qualitative classification of aphasia types (Fraser et al., 2013b; Fraser et al., 2013a; Peintner et al., 2008; Fraser et al., 2014c; Fraser et al., 2016; Vincze et al., 2016; Bucks et al., 2000; Guinn and Habash, 2012; Meilán et al., 2014; Jarrold et al., 2014) has been the primary focus of computational research into aphasia, whether in differentiating PWA’s and controls or between aphasia sub-types. Traditional features sets include features that target dysfluency, lexical diversity, syntactic deviation, and language complexity. Quantitative prediction methods focus on assessing speech-based features quantitatively with the goal of providing feedback to aphasic patients. Automatic speech recognition (ASR) systems developed for aphasic speech are used to automatically extract and align a number of feature sets (Le et al., 2018; Le et al., 2014), targeting specific suggested characteristics of Aphasia. In quantitative prediction, regression models are trained on the extracted features from a subset of the annotated aphasia data.

Information-theoretic approaches (Pakhomov et al., 2010) of using the perplexity of a trained language model have been investigated in the classification of aphasia types related to dementia. The primary contribution of this research is an n-gram statistical language model trained on speech from a healthy population and used to capture unusual words and sequences from the speech of patients with frontotemporal lobar degeneration (FTLD). This model was then used to measure the dissimilarity and degree of deviation from the healthy speech data, and found that the perplexity of a language model is sensitive to the semantic deficits in FTLD patients’ speech, which is often syntactically intact but is full of statistically unexpected word sequences. The perplexity index also discriminated mild from moderate-to-severely impaired FTLD patients, meaning that it is likewise sensitive to the severity of the aphasia. Few works, to our knowledge, attempt to model comprehension. Prud’hommeaux and Roark (2015), however, explore features based on the idea that non-aphasic individuals recounting a narrative are likely to use similar words and semantic concepts to the ones used in the narrative, and suggest that this similarity can be measured using techniques such as latent semantic analysis (LSA) or cosine distance. A key element in extracting instances of comprehension impairment is the assumption that breakdowns of

language understanding within conversation result in unexpected responses to a given comment or question. As outlined by Chinaei et al. (2017), these unexpected responses may follow certain trends, such as lack of continuation of topic or requests for repetition. In Watson (1999), those with Alzheimer’s Disease (AD) were most likely to respond during comprehension difficulties by either a lack of continuation (no contribution or elaboration on the topic, or complete change of topic) or reprise with dysfluency (a partial or complete repetition of the question with frequent pauses and filler words). This is in contrast to those without AD, who showed more preference for specific request for information or hypothesis formation (guessing missed information).

## 3. Data

Datasets containing various types of conversational language are available for use in training the methods within this approach. The main requirements being that they have a clear distinction between speakers and have some sort of turn-taking conversational flow. Effort was made to collect datasets of predominantly North American English, as the test set contains mainly North American participants or at least consists primarily of participants born in the United States. For our purposes, two datasets were source to be used separately: a dataset of aphasic language to be used as a test set with both Aphasic participants and controls (AphasiaBank) on which we can assess the extracted features, and a large non-aphasic corpus that can be used to generate statistical information and examples of presumed healthy speech (Reddit).

### 3.1. AphasiaBank

The primary aphasic corpus used in this research is AphasiaBank, a multimedia dataset of interactions between patients with aphasia (PWA) and research investigators, for the study of communication in aphasia (MacWhinney et al., 2011; Forbes et al., 2012). The data is collected by various research groups under varying conditions following these protocols. The basic structure of these protocols involves the research investigator asking open-ended questions to elicit spontaneous verbal responses from the patient. For example, the main AphasiaBank protocol contains questions such as “*How do you think your speech is these days?*” and “*Tell me as much of the story of Cinderella as you can*”. Alternatively, there is the Scripts protocol, which is less frequent, but is used by a small subset of the data (Le, 2017). The protocols contain four different discourse tasks, such as giving personal narratives in response to questions, picture descriptions, story telling, and procedural discourse. For these activities, investigators follow a script, which includes a second level prompt if the patient does not respond in ten seconds and an additional troubleshooting script with simplified questions if the patient is still not able to respond. The AphasiaBank dataset contains a total of 431 (255 Male, 176 Female) aphasic subjects and 214 (94 Male, 120 Female) controls, with an average age of 62.4 for the aphasic group and 58.9 for the control group. The distribution of diagnosed aphasia types is outlined in Table 1.

Aphasia Type	Gender		Total
Broca	66M	33F	99
Transmotor	5M	5F	10
Global	4M	0F	4
Wernicke	21M	9F	30
Conduction	40M	26F	66
Anomic	79M	60F	139
TransSensory	0M	2F	2
AphasicNoDiagnosis	28M	18F	46
NotAphasicByWAB	12M	23F	35

Table 1: Number of AphasiaBank participants for each type of Aphasia as classified by WAB-R AQ

Speech in AphasiaBank is transcribed using the CHAT format (MacWhinney, 2000), which includes annotation of filler words, repetition, non-verbal actions, and phonetic transcription in the International Phonetic Alphabet (IPA) of word-level errors. For the purpose of this work, annotations denoting auditory occurrences and physical movements are not retained. The text for both investigator comments and replies is pre-processed and normalized using the following procedure, where the text is first lower-cased, all non-alphabetic characters and punctuation are removed, and any paraphasias or neologisms marked in the annotation are replaced with an `<UNK>` token. Annotated speech segments between researcher and patient are extracted to create comment-reply pairs. To extract consistent comment-reply pairs, utterances that have been split in the original data, and thus do not have a direct pair with a question or comment for an investigator, are appended to the end of the preceding utterance. The resulting textual data consists of 18,038 comment-reply pairs for the aphasia subset and a total of 448,337 words (8439 unique words) of annotated aphasic speech. The control group includes an additional 2620 comment-reply pairs, with 354,620 total words and 10,012 unique words.

### 3.1.1. Participant-level Assessment Statistics

The AphasiaBank data provides additional information about the participant, such as a number of test scores that aim to assess the severity and type of aphasia of each aphasic speaker. This includes the Western Aphasia Battery-Revised (WAB-R) Aphasia Quotient (AQ) (Kertesz, 2006), which is the most useful for this research. WAB-R AQ is the most widely administered test in the AphasiaBank database, and is composed of multiple standardized sub-tests that targets specific aphasia-related impairments. WAB-R AQ has been shown to be a relatively reliable assessment of aphasia severity, with it demonstrating a high retest reliability in studies of chronic aphasia patients (Kertesz and Poole, 1974). Weighted performance over a number of sub-tests produces an overall score ranging from 0 to 100, that measures a speaker’s general linguistic abilities and severity of their aphasia (Kertesz, 2006).

The specific sub-test groups that WAB-R AQ is composed of include: Spontaneous Speech, Repetition, Naming/Word Finding, and Auditory-Verbal Comprehension. Scores over

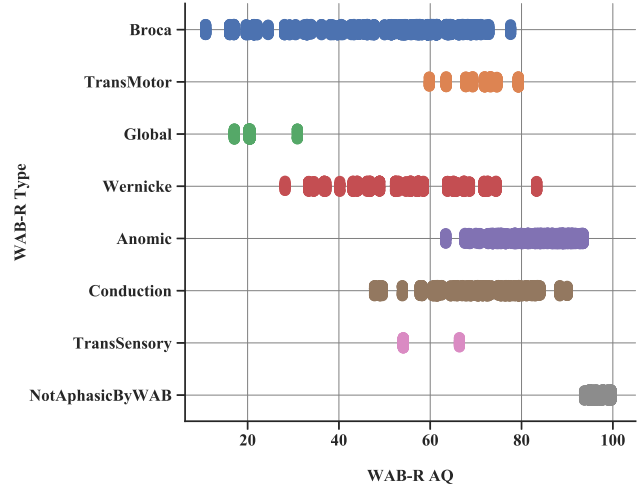


Figure 1: Distribution of WAB-R AQ Scores for each Aphasia Type (WAB-R AQ Type).

76 can be roughly classified as mild, whereas scores below 50 and 25 can be considered as severe and very severe respectively (Le, 2017). Following the WAB-AQ scores, the distribution of aphasia severity in the data is 47.0% mild, 38.6% moderate, 10.4% severe, and 3.9% very severe. The distribution of aphasia severity assessed by WAB-R AQ for each aphasia type is presented in Figure 1. For the purpose of evaluating fluent aphasia predictions, we consider the complete WAB-R AQ, as well as the Auditory-Verbal Comprehension sub-test scores.

The WAB-R Auditory-Verbal Comprehension sub-test scores offer the opportunity for us to evaluate our features on whether they do accurately capture information regarding comprehension impairments and not just additional information associated with other deficits related to aphasia. Auditory-Verbal Comprehension scores are assessed by yes/no questions that may be answered in either verbal or nonverbal fashion, word recognition tasks, and by response to sequential commands, with 10.0 being the upper bounds of the test. This score is aggregated with other sub-tests as a portion of the complete WAB-R AQ. In our data, Auditory-Verbal Comprehension scores exist for 351 speakers.

### 3.2. Reddit

Reddit is a social news aggregation, web content rating, and discussion website with over 234 million unique users, as of March 2019. The website is primarily in English, being the 6th most visited website in the United States, and with 53.9% of its users residing in the United States, and an additional 14.5% of its user base coming from the United Kingdom and Canada (Alexa Internet, 2018). Online community forums offer an abundant source of diverse structured semi-conversational textual data to be used for training, with Reddit’s being particularly easy to obtain. It should be considered semi-conversational due to the narrative quality of some comments, but there is a general assumption that threads are conversational in nature. Threads of comments are also divided hierarchically, so extracting comment relationship is possible. Though all datapoints cannot con-

firmed to be neurotypical or non-aphasic, the size of the dataset should minimize the impact of such outliers.

The bulk of Reddit comments dating back to its creation are obtained in JSON format from a repository prepared by (Baumgartner, 2018; Gaffney and Matias, 2018). Due to the size of the data, we only use a subset of the Reddit data. The data was naturally divided by subreddits, so a single subreddit with a still sizeable amount of data was chosen, r/IAmA (subreddits are denoted on Reddit using an r/ construction). The dialogue from this subreddit is generally representative of average healthy speech, as it is relatively serious in content, non-technical, and conversational. For normalization purposes, formatting tags are removed and double quotation marks were changed to single quotations. Links contain little relevant information for our purposes, so they were removed, along with the comments marked [DELETED] or [REMOVED]. To generate reasonable response lengths for conversation, comments longer than 50 words or 1000 characters were filtered, in addition to the removal of potential spam comments (with a user-assigned comment score  $\leq 1$ ). The data was further normalized to be better comparable to the other datasets used in this work. This included lowercasing the text, removing punctuation, and removing any commented links or quotations of other comments. The resulting dataset contains 1,050,699 sentence pairs, comprising of 768,348 unique words. The average number of tokens in a comment is roughly 16, where a comment is sometimes composed of multiple sentences.

## 4. Methods

Quantifiable measures of characteristic features of the fluent aphasia sub-type may be required to better accurately predict a general quantitative measure of aphasia severity. We propose multiple methods of extracting these measures, based on extensions of existing approaches in parallel domains, as well as additional novel approaches. We focus specifically on methods that extract features related to the production and comprehension issues found in fluent aphasic language.

### 4.1. Production Analysis Measures

This group of features targets aspects of fluent aphasia related to the production of language, such as sentence predictability and flow, along with occurrence of likely paraphasias or neologisms.

#### 4.1.1. Bigram Perplexity

In previous research introduced by (Pakhomov et al., 2010), bigram language model perplexity, as well as the out-of-vocabulary (OOV) rate, of utterances were shown to have a moderate best correlation ( $r=0.52$ ) with aphasia severity in dementia patients. For this approach, we suggest investigating whether we can extrapolate this approach for use with post-stroke PWAs and whether the same degree of correlation can be achieved. Following this research, we compute the probability of a sequence of words based on our non-aphasic Reddit corpus,  $P(W) = P(w_1, \dots, w_n)$ . To compute this, we want to consider the probability of a word given its previous context,  $P(w_n | w_{n-1})$ . The probability

for each bigram in our language model is computed as follows, where add-alpha smoothing is chosen and alpha  $\alpha$  is set to 0.02, to penalize OOV words.

$$P^*(w_n | w_{n-1}) = \frac{C(w_{n-1}, w_n) + \alpha}{C(w_{n-1}) + \alpha|V|}$$

Perplexity is then calculated for each utterance provided by the speaker, and the sum of all perplexity scores provides a speaker-level score. Perplexity in this case is measuring how well the given utterance mimics healthy speech when it comes to constructing probable strings of words.

#### 4.1.2. Out-of-vocabulary Rate

Out-of-vocabulary (OOV) rate may reflect the rate of paraphasia or neologisms in an utterance, with neologisms in particular being characteristic in some fluent aphasias, such as Wernicke’s. Often seen in the speech of patients with fluent aphasia are utterances that are long, but full of such neologisms. Therefore, a vocabulary is selected based on our non-aphasic corpus, and the target calculation would be the sum of all words not found in vocabulary over the total words in an utterance.

#### 4.1.3. Text Imputation Similarity

Despite sounding fluent at the surface level, fluent aphasia speech often lacks semantic cohesion within an utterance. Words selected by aphasic individuals may appear semantically incongruous with other nearby words in the utterance, although some meaning may still be parsed from the utterance.

A solution to capture this aspect would be to use a language model with a much greater ngram size. However, this would require a huge corpus and rare but semantically plausible utterances would be unfairly penalized by the language model. Word embeddings in this case give much more flexibility. To describe this approach, we will consider  $N$  be the length of the input sentence, and  $n = 0$  the index of the current word in the sentence. The process can then be summarised into the following steps, as shown in Figure 2, assuming a sentence string as input:

Given an utterance string  $S$ , the string is tokenized, such that  $S = \{w_1, \dots, w_N\}$ .  $N$  copies of the input strings are created, where for each string, the  $n_{th} + 1$  word is masked with the [MASK] token. Then, each masked word is predicted from the complete sentence context and resulting predicted words are concatenate to produce an output string  $O$ . The cosine similarity between the sentence vector of the original input utterance  $v_S$  and the sentence vector of the output string  $v_O$  is then compared.

### 4.2. Comprehension Analysis Measures

Comprehension analysis measures uniquely target fluent aphasia by assessing response predictability and sudden changes in topic, where unpredictable responses may signify a lapse in comprehension.

#### 4.2.1. Question-Answer Similarity

Semantic relation between questions or statements and their responses are of particular interest, due to the proposed hypothesis that responses denoting an error in understanding

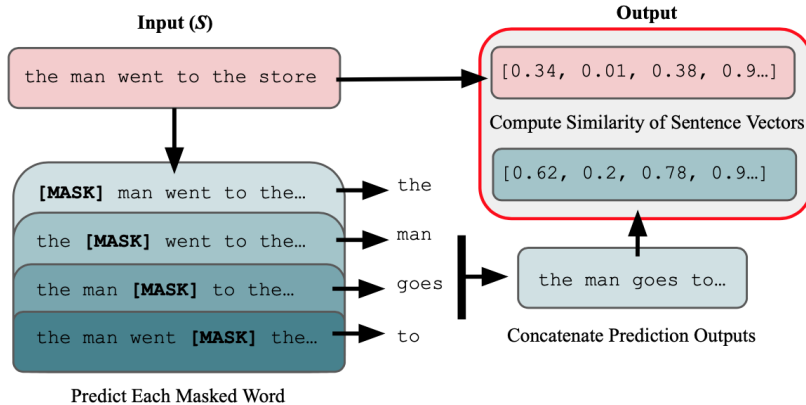


Figure 2: Text Imputation feature extraction process.

will be dissimilar to the question in words-use and semantics. To capture this, a basic measure of the cosine similarity between the sentence representations of the questions and answers in a dataset of aphasic speech can be obtained. For each question-answer pair, sentence representations of the question and answer is separately produced. The cosine similarity between the two vectors will then be computed to produce a score. Our hypothesis will be that lower similarity between the two sentence vectors will indicate less semantic overlap between the content of the sentences, meaning that the response in the question-answer pair may not be semantically coherent with the question.

For example, given a question such as 'What did you see at the zoo?' or 'What's your favourite animal?', answers containing few or no words related to zoo or animal, may indicate a misunderstanding of the question. The use of good sentence representations from word embedding models is especially useful in this task, because given our examples, a favourite animal might be uncommon, but still semantically related to animal.

#### 4.2.2. Closest Question-Answer Pair

An expected and appropriate answer to a given question is assumed to closely resemble other appropriate answers to the same or similar questions. By finding the most similar question match to the question portion of a question-answer pair within a corpus of healthy speech, the question match's corresponding answer can then be compared to the answer in the input question-answer pair. Demonstrated in Figure 3, this is done by first generating the sentence representations of the input question and answer (from AphasiaBank, in our case), as well as all questions and answers in the healthy corpus (Reddit). Then, given the input question-answer sentence pair  $s_{q,a}$  and a non-aphasic speech corpus of question-answer sentence pairs  $C = \{c_{1,1}, \dots, c_{q,a}\}$ , where  $q = a$ . Sentence vectors for  $s_q$  and  $s_a$  are generated. For each sentence pair in  $C$ , the vector representation for  $c_q$  is also generated, resulting in a set of corpus sentence vectors  $VQ$  of length  $(C)$ . For each vector in  $VQ$ , its cosine similarity with  $s_q$  is computed. Selecting the vector  $VQ_q$  with greatest similarity with  $s_q$ , the sentence representation of  $c_a$  is retrieved. Finally, the cosine similarity between

the vectors of  $s_a$  and  $c_a$  is computed as the feature for this approach..

#### 4.2.3. Binary Sentence Pair Classification

We leverage a binary classification approach using Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) to predict the likelihood of a given sentence pair being related. Our assumption is that question-answer sentence pairs that are predicted to be related based on our non-aphasic corpus are likely to contain semantically coherent answers to the questions, and are therefore unlikely to be characterized as a misunderstanding.

To train this approach, we first gather the non-aphasic corpora question and answer pairs collected from the Reddit dataset as positive samples and artificially fabricate negative sample pairs, by randomly sampling accompanying answers segments for each question segment from the corpus. This gives us a training sets of sentence pairs double the size of the non-aphasic corpus. With this new training set, we fine-tune a sentence pair classifier with two output classes, whether the sentences contains a valid question and answer pair or not.

The sentence pair classifier functions using the pre-trained BERT model, *bert-base-uncased*, with an additional attached classification layer. The original BERT model includes layers for language model decoding and classification, but these are not used in fine-tuning the sentence pair classifier. The sentence pair classifier uses the base model to encode the sentence representations, followed by an additional hidden, non-linear layer and the classification layer. Because the classifier uses BERT to encode the sentence representations, to fine-tune, the training data must be structured the way BERT expects, with an initial [CLS] token at the beginning of every sequence (question-answer pair), necessary for classification with BERT, and a [SEP] token between the two sentences. The classifier is then given the target question-answer pairs to generate probabilities for the two classes. The probability of the second class, which is the probability of the two sentences being a pair, is used as an aphasia severity feature.

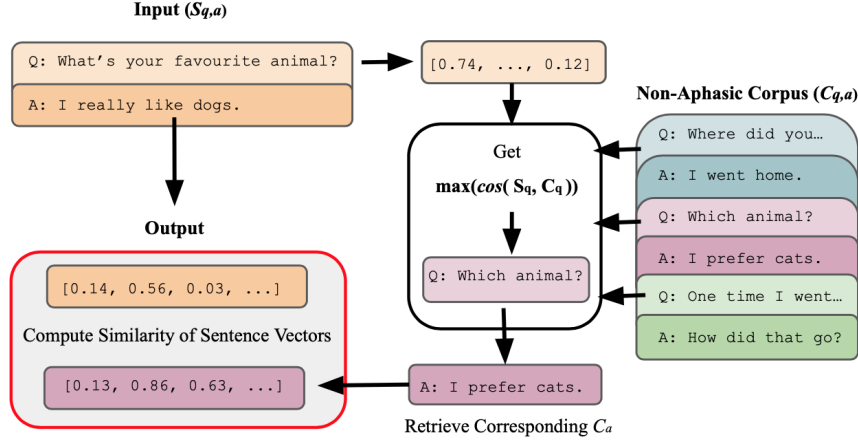


Figure 3: Closest Match feature extraction process.

## 5. Experimental Setup

To allow for easy comparison and combination of features that may have wildly different relationships with the data, we z-normalized all extracted features based on statistics from the control participants of AphasiaBank. Z-normalization produces a standard score useful for speaker comparison against the control group, and is calculated by subtracting the control population mean from each individual computed score and then dividing the difference by the standard deviation of the control group. With the produced feature sets, organized into groups, the goal is to produce a measure from a sample of aphasic speech that aligns with the speaker’s manually diagnosed score of aphasia severity. We select only aphasic speakers who have been assigned the aphasia severity score of interest in the AphasiaBank data.

To model aphasia severity with the grouped feature sets, we use Linear Regression, Support Vector Regression (SVR), and Random Forest Regression (RFR) implemented with Scikit-learn (Pedregosa et al., 2011). The models are trained for both WAB-R AQ and Sentence Comprehension score prediction, and Pearson Correlation between the predicted results of the test set and the target scores is used to evaluate the model. The data is split at the speaker-level using four fold cross-validation, where one fourth of the data is held out as a test set during each fold, and the remaining fourths are used for training the model. While the features themselves do not require annotated aphasic data to extract, to utilize the multiple features in the most optimal way, some amount of annotated and scored aphasic data is required to fit the prediction model. We, however, also report the individual features strengths in our results. Hyperparameter selection using 10-fold cross-validation is preformed prior to training, using the GridSearchCV function in Scikit-learn. For each model, the hyperparameters tested were:

**Linear Regression** Intercept  $\{True, False\}$ , and if intercept is calculated, then normalize  $\{True, False\}$ .

**Support Vector Regression** Penalty term  $C$

$\{1.0, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$ , slack parameter  $\epsilon \in \{1.0, 10^{-1}, 10^{-2}, 10^{-3}\}$ , kernel type  $\{rbf, linear\}$ , and shrinking heuristic  $\{True, False\}$ .

**Random Forest Regression** Number of trees  $\{10, 100, 200\}$ , function use to measure quality of split  $\{mse, mae\}$ , and the max number of features to consider  $\{auto, sqrt\}$ .

## 6. Results

### 6.1. Individual Feature Correlations

For all feature correlations, due to likely monotonic relationships once the control data is added, we first compare the Spearman correlations of features for a combined set of aphasic and AphasiaBank control participants, in addition to the aphasic participants evaluation set, holding out the control data. Table 2 presents the correlations of all proposed features. As mentioned previously, since the control group is not naturally given a WAB-R AQ score, the scores for this group were automatically set to the upper limits of the WAB-R AQ, which is 100.0 for the full score, and 10.0 for the Auditory-Verbal Comprehension component. A feature that has a high correlation in the aphasia-only set compared to the combined control/aphasic set, likely can distinguish between more nuanced aphasia severity levels and not just between healthy controls and person with aphasia. All features have a p-value less than 0.001 in their Spearman correlations.

The feature with the strongest correlation with WAB-R AQ without the control data was the Sentence Classifier with a correlation of 0.558. This holds true also for the comprehension scores, with a correlation of 0.415. The weakest feature for the no control data is then Bigram Perplexity, likewise for both WAB-R AQ and Comprehensions with a correlation of -0.335 and -0.228 respectively. Bigram Perplexity had a very weak Pearson correlation, but still has a moderate Spearman here, indicating that it may not perform well with continuous data, but could be a useful feature in classification tasks.

Table 2: Individual Spearman correlations for all proposed features

FEATURES	WAB-R AQ		Aud-Vbl Comprehension	
	With Control	No Control	With Control	No Control
	Bigram Perplexity	-0.469	-0.335	-0.382
OOV Rate	<b>-0.675</b>	-0.465	<b>-0.553</b>	-0.251
Text Imputation	0.505	0.372	0.395	0.273
QA Similarity	0.281	0.345	0.221	0.271
Closest QA Pair	0.472	0.406	0.372	0.321
Sentence Classifier	0.33	<b>0.558</b>	0.271	<b>0.415</b>

With control data added, OOV Rate has a relatively strong correlation of -0.675. This comes with an increase of 0.21 for WAB-R AQ and 0.302 for Comprehension, compared to its correlation with the non-control data, suggesting that it may be a particularly useful feature in distinguishing healthy and aphasic individuals. Bigram Perplexity, Text Imputation, and Closest QA Pair also found a increased correlation to WAB-R when control data was added.

The Sentence Classifier feature did not correlate well with the added control data for either evaluation set, with a 0.228 difference from the non-control data. This brings it from being the most correlated feature for the non-control data to the second least with the added control data. We are unsure why this is, though we hypothesize that it is capturing variation within the control group that is not represented due to the uniform scoring the controls received. QA Similarity also correlated more strongly without the control group, though not as drastically as the Sentence Classifier.

## 6.2. Quantitative Aphasia Severity Prediction

One of our primary goals is to predict aphasia severity. We attempted to do so by using the features we computed in a regression model. In our preliminary investigations we utilize the three regression models for comparison. This is done on the AphasiaBank aphasic dataset with the exclusion of the AphasiaBank control group.

We compare this with a baseline consisting of a high performing Lexical Diversity and Complexity feature set (LEX) previously used for this task (Le et al., 2018; Fraser et al., 2013b), which consists of Type-Token Ratio, a mapping of words and their frequencies in American English called the SUBTL norms (Brysbaert and New, 2009), and four additional Bristol norm word-level measures (Imageability, Age of Acquisition, Familiarity, and Phones), produced by the combined work of Stadthagen-Gonzalez and Davis (2006) and Gilhooly and Logie (1980). For our application of the baseline we achieved a Pearson correlation of 0.621 for predicting WAB-R AQ and 0.439 for predicting Auditory-Verbal Comprehension with the Support Vector Regression model. Random Forest Regression performed the best overall and for the baseline, with a correlation for 0.703 for WAB-R AQ prediction and 0.523 for Auditory-Verbal Comprehension.

We grouped our feature sets together into Production Analysis Measures (PROD), consisting of the Bigram Perplexity, OOV Rate, and Text Imputation features, and

Comprehension Analysis Measures (COMP), consisting of the Question-Answer Similarity, Closest Question-Answer Pair, and Binary Sentence Pair Classification features. The proposed feature sets in the WAB-R AQ prediction task alone do not achieved the same level of results as the baseline alone, with an average correlation across models of 0.434 for Production features and 0.574 for Comprehension features. Of course each of the proposed groups consist of half of the features as the Lexical feature set. The Linear model is an exception, however, as it performs unexpectedly well with the Comprehension feature set, beating the baseline with the Comprehension features alone. Over the three models, the best performing set of features for this task is the combined baseline and the comprehension features (LEX + COMP), which given us an average correlation of 0.692 and an improvement over the baseline of 0.066. It is also the best performing feature set for both the Linear model and SVR, with Random Forest Regression performing best with all features (LEX + PROD + COMP). The Linear model performed overall, surprisingly well for the task, yielding a slightly stronger correlation than Support Vector Regression.

Predictions for Auditory-Verbal Comprehension scores follow a similar pattern to the WAB-R AQ task. The Lexical and Comprehension feature set (LEX + COMP) prediction correlations remain the best performing with an average correlation of 0.490 and an improvement over the baseline of 0.037. In the prediction of Comprehension scores, the Comprehension feature set generally performed more closely to the baseline than in the WAB-R task, whereas the Production feature set performed equally as poorly compared to the baseline as it did in predicting WAB-R AQ.

It is interesting to note that any inclusion of the Production feature set in both tasks worsened the performance of the model, with the exception of Random Forest Regression, which had the best results with the Lexical and Production features sets (LEX + PROD). This suggests that some sort of feature selection may need to be applied.

### 6.2.1. Feature Selection

Certain particularities stand out in the model prediction results which leaves additional consideration to the efficacy of some features, such as the decrease in improvement following the addition of the Production feature set (or Comprehension feature sets for Random Forest Regression) and the poor linear correlations of some features. For this reason, we apply a feature selection method to the data.

Table 3: Prediction model results for 3 feature sets after applying feature selection, on the two evaluations sets: WAB-R AQ and Auditory-Verbal (Aud-Vbl) Comprehension.

Feature Sets	Pearson $r$ (p-value)		
	Linear	Support Vector	Random Forest
WAB-R Baseline	0.563	0.617	0.691
WAB-R Proposed	0.616	0.619	0.576
WAB-R Combined	0.715	0.714	0.74
Aud-Vbl Baseline	0.403	0.428	0.494
Aud-Vbl Proposed	0.414 (0.001)	0.423	0.321 (0.01)
Aud-Vbl Combined	0.488	0.491	0.537

We apply the Boruta algorithm, using `Boruta.py` and `Scikit-learn`, to optimize prediction results and as an easily interpretable method for feature selection. With this we can determine which combination of features yield the best performance from our models. The Boruta algorithm (Kursa et al., 2010) is a recursive feature elimination method. It functions by adding randomness to the data in creating shuffled copies of all the features. Then we give this extended feature set to be fit to the evaluation data using a Random Forest Regressor. Feature importance is measured during training of the regressor, using Mean Decrease Accuracy, where higher means indicate more importance. For each iteration of training, the algorithm checks if a feature has a higher importance than the best of its shuffled copies and removes features it deems as unimportant.

For each evaluation measure (WAB-R AQ Aud-Vbl Comprehension), we run feature selection on three sets of the features, one including the only baseline Lexical features, one with only our proposed feature, and one with all features. We report the model results in Table 3. P-values for all feature set predictions were less than 0.001, unless otherwise specified. For WAB-R AQ prediction, the following proposed features were selected: OOV Rate, Text Imputation, Closest QA Pair, Sentence Classifier

For Auditory-Verbal Comprehension prediction, only Closest QA Pair and the Binary Sentence Classifier probabilities were selected, both with the baseline features and without. For Auditory-Verbal Comprehension, production based features in particular were excluded during selection, such as Phone Length in the Lexical features set, and OOV Rate. In all cases the Combined Baseline and Proposed feature set performed best, though the results using proposed selected features alone correlated better than the baseline feature in all models except Random Forest Regression for WAB-R AQ Prediction. On the other hand, Random Forest Regression provided the best results using all features for both WAB-R AQ and Auditory-Verbal Comprehension score prediction.

## 7. Conclusion

In this work, we proposed methods for extracting six features we hypothesized would be useful in modelling symptoms consequent of fluent aphasia, such as comprehension impairments, semantic incoherence, and increased likelihood of paraphasias and neologisms. We make primary

use of word and sentence representation to better assess these aspects. Our chosen approach utilized the perceived dissimilarity between aphasic and non-aphasic speech and thus did not require any annotated data of aphasic speech to obtain the proposed features. We assess the performance of our features by investigating how they benefit the task of quantitative aphasia severity prediction. Framing the task as a regression problem, and given a set of data with manually assigned aphasia severity scores, we evaluated the linear correlation of the predicted scores using our proposed features against the gold-standard severity scores. We compared these results to a baseline based on work by Le et al. (2018; Fraser et al. (2013b)). Most of the proposed features alone were found to have moderate correlation with the evaluation scores, and after applying feature selection, the proposed features performed better or equal to the baseline in the regression task using Linear Regression and Support Vector Regression. For all regression models, the combined baseline and proposed features yielded the best results in all evaluation cases. Specifically, we found that the task benefits most from the inclusion of BERT sentence representations fine-tuned on a large amount of conversational data.

This work has also raised a number of questions and possible avenues for future work in this research area. Since scores were predicted at the utterance-level and then averaged, a wider range of statistics for the proposed features may yield better results, as was previously investigated by (Le et al., 2018). Likewise, given a larger dataset of aphasic language for each aphasia subtype, variations between subtypes could offer further structured results that highlight the difference between fluent and non-fluent aphasia. The practical applications of such a task using more robust feature sets, automatic speech recognition, and utterance-level assessment is also worth consideration.

## 8. Acknowledgements

This work was partially funded by the European Union’s Horizon 2020 grant agreement No. 777107 (Precise4Q) and by the BMBF project DeepLee ((01IW17001).

## 9. Bibliographical References

Adams, J., Bedrick, S., Fergadiotis, G., Gorman, K., and van Santen, J. (2017). Target word prediction and para-

- phasia classification in spoken discourse. In *BioNLP 2017*, pages 1–8.
- Alexa Internet. (2018). Reddit.com site info. [Online; accessed 2018-02-28].
- Baumgartner, J. (2018). [dataset] subreddit data for over 750,000 reddit subreddits. [Online; accessed 2019-05-21].
- Bhogal, S. K., Teasell, R., Speechley, M., and Albert, M. (2003). Intensity of aphasia therapy, impact on recovery. *Stroke-a Journal of Cerebral Circulation*, 34(4):987–991.
- Brysbart, M. and New, B. (2009). Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior research methods*, 41(4):977–990.
- Bucks, R. S., Singh, S., Cuerden, J. M., and Wilcock, G. K. (2000). Analysis of spontaneous, conversational speech in dementia of alzheimer type: Evaluation of an objective technique for analysing lexical performance. *Aphasiology*, 14(1):71–91.
- Chinaei, H., Currie, L. C., Danks, A., Lin, H., Mehta, T., and Rudzicz, F. (2017). Identifying and avoiding confusion in dialogue with people with alzheimer’s disease. *Computational Linguistics*, 43(2):377–406.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Forbes, M. M., Fromm, D., and MacWhinney, B. (2012). Aphasiabank: A resource for clinicians. In *Seminars in speech and language*, volume 33, pages 217–222. Thieme Medical Publishers.
- Fraser, K., Rudzicz, F., Graham, N., and Rochon, E. (2013a). Automatic speech recognition in the diagnosis of primary progressive aphasia. In *Proceedings of the fourth workshop on speech and language processing for assistive technologies*, pages 47–54.
- Fraser, K. C., Rudzicz, F., and Rochon, E. (2013b). Using text and acoustic features to diagnose progressive aphasia and its subtypes. In *INTERSPEECH*, pages 2177–2181.
- Fraser, K. C., Hirst, G., Graham, N. L., Meltzer, J. A., Black, S. E., and Rochon, E. (2014a). Comparison of different feature sets for identification of variants in progressive aphasia. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 17–26.
- Fraser, K. C., Hirst, G., Meltzer, J. A., Mack, J. E., and Thompson, C. K. (2014b). Using statistical parsing to detect agrammatic aphasia. *Proceedings of BioNLP 2014*, pages 134–142.
- Fraser, K. C., Meltzer, J. A., Graham, N. L., Leonard, C., Hirst, G., Black, S. E., and Rochon, E. (2014c). Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. *cortex*, 55:43–60.
- 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.
- Gaffney, D. and Matias, J. N. (2018). Caveat emptor, computational social science: Large-scale missing data in a widely-published reddit corpus. *PloS one*, 13(7):e0200162.
- Gilhooly, K. J. and Logie, R. H. (1980). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior research methods & instrumentation*, 12(4):395–427.
- Guinn, C. I. and Habash, A. (2012). Language analysis of speakers with dementia of the alzheimer’s type. In *2012 AAAI Fall Symposium Series*.
- Jarrold, W., Peintner, B., Wilkins, D., Vergry, D., Richey, C., Gorno-Tempini, M. L., and Ogar, J. (2014). Aided diagnosis of dementia type through computer-based analysis of spontaneous speech. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 27–37.
- Kertesz, A. and Poole, E. (1974). The aphasia quotient: the taxonomic approach to measurement of aphasic disability. *Canadian Journal of Neurological Sciences*, 1(1):7–16.
- Kertesz, A. (2006). Western aphasia battery-revised (wab-r): Harcourt assessment. *San Antonio, TX*.
- Kursa, M. B., Jankowski, A., and Rudnicki, W. R. (2010). Boruta - a system for feature selection. *Fundam. Inf.*, 101(4):271–285, December.
- Laska, A., Kahan, T., Hellblom, A., Murray, V., and Von Arbin, M. (2011). A randomized controlled trial on very early speech and language therapy in acute stroke patients with aphasia. *Cerebrovascular diseases extra*, 1(1):66–74.
- Le, D., Licata, K., Mercado, E., Persad, C., and Provost, E. M. (2014). Automatic analysis of speech quality for aphasia treatment. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4853–4857. IEEE.
- Le, D., Licata, K., and Provost, E. M. (2017). Automatic paraphasia detection from aphasic speech: A preliminary study. In *Interspeech*, pages 294–298.
- Le, D., Licata, K., and Provost, E. M. (2018). Automatic quantitative analysis of spontaneous aphasic speech. *Speech Communication*, 100:1–12.
- Le, D. (2017). *Towards Automatic Speech-Language Assessment for Aphasia Rehabilitation*. Ph.D. thesis.
- MacWhinney, B., Fromm, D., Forbes, M., and Holland, A. (2011). Aphasiabank: Methods for studying discourse. *Aphasiology*, 25(11):1286–1307.
- MacWhinney, B. (2000). The chldes project: Tools for analyzing talk: Volume i: Transcription format and programs, volume ii: The database.
- Meilán, J. J. G., Martínez-Sánchez, F., Carro, J., López, D. E., Millian-Morell, L., and Arana, J. M. (2014). Speech in alzheimer’s disease: Can temporal and acoustic parameters discriminate dementia? *Dementia and Geriatric Cognitive Disorders*, 37(5-6):327–334.
- Pakhomov, S. V., Smith, G. E., Marino, S., Birnbaum, A.,

- Graff-Radford, N., Caselli, R., Boeve, B., and Knopman, D. S. (2010). A computerized technique to assess language use patterns in patients with frontotemporal dementia. *Journal of neurolinguistics*, 23(2):127–144.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Peintner, B., Jarrold, W., Vergyri, D., Richey, C., Tempini, M. L. G., and Ogar, J. (2008). Learning diagnostic models using speech and language measures. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4648–4651. IEEE.
- Prud'hommeaux, E. and Roark, B. (2015). Graph-based word alignment for clinical language evaluation. *Computational Linguistics*, 41(4):549–578.
- Stadhagen-Gonzalez, H. and Davis, C. J. (2006). The bristol norms for age of acquisition, imageability, and familiarity. *Behavior research methods*, 38(4):598–605.
- Thomas, C., Keselj, V., Cercone, N., Rockwood, K., and Asp, E. (2005). Automatic detection and rating of dementia of alzheimer type through lexical analysis of spontaneous speech. In *IEEE International Conference Mechatronics and Automation, 2005*, volume 3, pages 1569–1574. IEEE.
- Vincze, V., Gosztolya, G., Tóth, L., Hoffmann, I., Szatlóczki, G., Bánréti, Z., Pákáski, M., and Kálmán, J. (2016). Detecting mild cognitive impairment by exploiting linguistic information from transcripts. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 181–187.
- Watson, C. M. (1999). An analysis of trouble and repair in the natural conversations of people with dementia of the alzheimer's type. *Aphasiology*, 13(3):195–218.

# Linguistic Markers of Anorexia Nervosa: Preliminary Data from a Prospective Observational Study

Giulia Minori<sup>1</sup>, Gloria Gagliardi<sup>2</sup>, Vittoria Cuteri<sup>3,4</sup>, Fabio Tamburini<sup>1</sup>,  
Elisabetta Malaspina<sup>4</sup>, Paola Gualandi<sup>4</sup>, Francesca Rossi<sup>4</sup>,  
Filomena Moscano<sup>4</sup>, Valentina Francia<sup>4</sup>, Antonia Parmeggiani<sup>3,4</sup>

<sup>1</sup> Department of Classical Philology and Italian Studies, University of Bologna, Italy

<sup>2</sup> Department of Literary, Linguistics and Comparative Studies - University of Naples "L'Orientale", Italy

<sup>3</sup> Department of Medical and Surgical Sciences - University of Bologna, Italy

<sup>4</sup> Regional Center of Eating Disorders, Child Neuropsychiatry Unit - Policlinico Sant'Orsola Malpighi, Bologna, Italy

giulia.minori@studio.unibo.it, ggagliardi@unior.it, cuteri.vittoria@gmail.com, fabio.tamburini@unibo.it  
elisabetta.malaspina@aosp.bo.it, paola.gualandi@aosp.bo.it, francesca\_rossi@aosp.bo.it,  
filomena.moscano@aosp.bo.it, valentina.francia@aosp.bo.it, antonia.parmeggiani@unibo.it

## Abstract

Recent works indicated the potential relevance of Natural Language Processing techniques for the detection of clinical conditions. This paper tries to address the issue in the Eating Disorder domain, by exploiting "linguistic biomarkers" for Anorexia Nervosa (AN) detection in female teenagers. We hypothesize that (i) disturbances in self-perceived body image, black and white thinking and mood changes strongly associated with AN disorder can result in altered linguistic patterns; and (ii) these subtle modifications can be identified by means of NLP tools, acting as early proxy measures for the disorder. To this aim, we enrolled 51 participants (age range: 14-18): 17 girls with a clinical diagnosis of Anorexia Nervosa and 34 normal weighted peers, matched by gender, age and educational level. Both the groups were asked to produce three written texts (around 10-15 lines long), i.e. two autobiographical narratives and a short description of a complex figure. A rich set of linguistic features was extracted from the text samples and the statistical significance in pinpointing the pathological process was measured. Our preliminary results show that subtle language disruptions, mainly at the lexical and syntactic level, can actually represent an early but reliable marker of the disease. However, an analysis on a bigger cohort with follow-up information, still ongoing, is needed to consolidate this assumption.

**Keywords:** Linguistic Markers, Feeding and Eating Disorders, Anorexia Nervosa

## 1. Background

### 1.1 Feeding and Eating Disorders: the case of Anorexia Nervosa

According to DSM-5 definition (American Psychiatric Association, 2013), Feeding and Eating Disorders (FED) are characterized by "a persistent disturbance of eating or eating-related behavior that results in the altered consumption or absorption of food and that significantly impairs physical health or psychosocial functioning".

Among these clinical conditions, Anorexia Nervosa (ICD-10-CM codes: F50.01 and F50.02 (World Health Organization, 1993; World Health Organization, 1995)) takes on a special relevance, due to both epidemiological reasons and medical outcomes. As a matter of fact, AN is relatively common among young women:<sup>1</sup> although community studies assessing the incidence of eating disorders are scarce, one-year prevalence rate of AN has been calculated as 370 per 100 000 young females (Hoek, 1993; Smink et al., 2012). The majority of AN patients in the community do not enter the mental healthcare system. All eating disorders have an elevated mortality risk; however, AN is the most striking disease, showing the highest mortality rates among psychiatric pathologies, 5.1 deaths per 1000 person-years, of which 1.3 deaths resulted from suicide (Harris and Barraclough, 1998; Arcelus et al., 2011).

There are three essential diagnostic features of AN (American Psychiatric Association, 2013): (i) persistent energy

intake restriction, leading to a significantly low body weight (i.e., less than minimally normal or, for children and adolescents, less than that minimally expected) in the context of age, sex, developmental trajectory, and physical health; (ii) intense fear of gaining weight or of becoming fat (also known as "fat phobia"), or persistent behavior that interferes with weight gain, usually not alleviated by slimming; and (iii) a disturbance in self-perceived weight or shape.

Body mass index (BMI; calculated as weight in kilograms/height in meters<sup>2</sup>) is the common measure to assess criterion (i). For adults, a BMI of 18.5 kg/m<sup>2</sup> has been employed by the World Health Organization (WHO) as the threshold of normal body weight (Cole et al., 2007). From a psychological point of view, weight loss is often viewed by AN patients as a sign of extraordinary self-discipline, whereas weight gain is perceived as an unacceptable failure. Inflexible thinking is a core feature of the disorder, as well as narrow, rigid behaviour, almost disconnected from the somatic experience. Although some AN individuals may acknowledge being thin, they often do not recognize the serious medical consequences of their serious malnourished state; they either lack insight into or deny the problem.

A prompt identification (and treatment) of symptoms is linked to better outcomes (Herzog et al., 1996). Unfortunately, the diagnosis of AN is often elusive, and more than one half of all cases go undetected in the primary care setting (Becker et al., 1999). Therefore, current

<sup>1</sup> AN is far less common in males, with clinical populations generally reflecting approximately a 10:1 female-to-male ratio (American Psychiatric Association, 2013).

research continues to emphasise the need for novel reliable strategies in order to identify even early warning signs.

## 1.2 Linguistics and Natural Language Processing for the medical science: a growing area of study

Over the last few years, a growing body of linguistic studies have been devoted to speech and language disorders and remediation. This fairly new branch of linguistics, called “Clinical Linguistics” (Crystal, 1981), is constructing outline sketches of communicative disabilities, supporting the work of speech and language therapists and neuropsychologists. Within this context, a number of works have been published on “linguistic profiles” of various clinical populations (Marini and Carlomagno, 2004; Adornetti, 2018; Gagliardi, 2019): for example, linguistic deficits (mainly at syntactic and pragmatic level) have been reported in several neurodegenerative diseases such as dementia (Boschi et al., 2017; Beltrami et al., 2018), where language disruption is a common finding both at the earliest stages and in full-blown pathology; alterations have been extensively described in scientific literature on dysphonia and dysarthria, especially in the hypokinetic forms resulting from damage to the basal ganglia (such as in Huntington's disease, Progressive Supranuclear Palsy or Parkinsonism (Gagnon et al., 2018; Catricalà et al., 2019; Altmann and Troche, 2011; Montemurro et al., 2019)); some studies dealt with the linguistic habits of psychopathologies, e.g. schizophrenia (Dovetto, 2015; Bambini et al., 2016), personality disorder (Arntz et al., 2012), anxiety and depression (Ramirez-Esparza et al., 2008; Brockmeyer et al., 2015; Bernard et al., 2016; Edwards and Holtzman, 2017; Zimmermann et al., 2017; Al-Mosaiwi and Johnstone, 2018; Smirnova et al., 2018). However, a very limited number of papers have been devoted to linguistic changes in patients with eating disorders (Lyons et al., 2006; Espeset et al., 2012; Skårderud, 2007a; Skårderud, 2007b; Wolf et al., 2013; Brockmeyer et al., 2013; Spinczyk et al., 2018).

Thanks to automated computational methods, progress in the field has been breathtaking. The development of new sophisticated techniques from Natural Language Processing (NLP) have been used to analyze written and spoken texts, revealing latent patterns and regularities of pathological languages.

This subtle language disruptions can be employed as “digital biomarkers”, namely objective, quantifiable behavioral data which can be collected and measured by means of digital devices, allowing for a low-cost pathology detection and classification.

Dementia assessment is a key domain of NLP application for medical science, coming up with relevant results (Vincze et al., 2016; Asgari et al., 2017; Beltrami et al., 2018; Tóth et al., 2018; Themistocleous et al., 2018; Gosztolya et al., 2019; Fraser et al., 2019a; Fraser et al., 2019b), but this approach is spreading rapidly through the community (Spinczyk et al., 2018; Trotzek et al., 2018).

## 1.3 Linguistic profile of Anorexia Nervosa: a brief sketch

Little research has addressed the linguistic profiles of AN: some interesting studies focused on differences in self-presentation written texts of individuals who publicly defend AN as a lifestyle (“pro-ana”) and individuals who identify themselves as recovering from anorexia; others

investigated body’s symbolic role in the course of illness and “concretized metaphors”, i.e. “instances where the metaphors are not experienced as indirect expressions showing something thus mediated, but they are experienced as direct and bodily revelations of a concrete reality” (Enckell, 2002; Skårderud, 2007a); in layman's terms, emotions are concretised.

With regard to the first point, pro-anorexics and recovering anorexics engage in distinct linguistic self-presentation styles: the analysis of linguistic cues of emotional processes revealed that pro-anorexics usually use more positive emotional words (e.g. “happy”, “good”), a lower rate of anxiety words (e.g. “afraid”, “scared”) and fewer cognitive mechanism words (specifically insight and causation words, e.g. “cause”, “realize”) than recovering anorexics (Lyons et al., 2006; Wolf et al., 2013). Moreover, pro-anorexics display lower levels of self-directed attention, since they make fewer first person singular self-references; their texts contain more present tense verbs and fewer past tense verbs, suggesting a focus on the present experience rather than on the past. With regard to the prevalence of AN-related psychological concerns, pro-anorexics were more preoccupied with eating (e.g. “food”, “meal”, “diet”) and less with school-related issues (e.g. “exam”, “study”) and death (e.g. “dead”, “death”, “coffin”).

Compared with recovery and control blogs, pro-eating disorder written productions contain an exceptionally high proportion of exclamation marks but much fewer question marks: according to (Wolf et al., 2013), this might reflect a form of complexity reduction at the syntactical level. Furthermore, exclamation marks are often used as an orthographic intensifier, indicating a strong self-affirmation (Rubin and Greene, 1992), whereas the infrequent use of question marks might indicate a reduced tendency to express insecurity and fears (Wolf et al., 2013). This strong self-focus enters into combination with a low social relatedness. Pro-ana bloggers appear to be less connected with the outside world and real-life relationships (Gavin et al., 2008): this tendency is further supported by a low third-person plural pronoun use.

Taken together, these observations are consistent with an interpretation of pro-anorexics’ language use as a coping strategy aimed at stabilizing them emotionally, experiencing a sense of control over the illness, namely a mechanism of self-defense.

With respect to the second point, the work of (Skårderud, 2007a; Skårderud, 2007b) addressed the striking clinical feature of concreteness of symptoms, due to body image fluctuation. Numerous sentences of AN texts instantiate symbolisation via the body: these physical metaphors show a striking closeness and a primary relation between emotions and different sensorimotor experiences (e.g. heaviness/lightness: “*I dream of being so light that I can float in the air. Then I can move down the main street among the people, one meter above the ground, and I will feel that all my worries are gone, lifted off my shoulders*”; “*I feel sad. And when I am sad, I feel burdened and heavy... and then comes the urge to lose weight*”).

Quoting the author, “these bodily metaphors do not function mainly as representations [...], but as presentations which are experienced as concrete facts here-and-now and are difficult to negotiate with. The ‘as-if’ quality of the more abstract meaning of the metaphor is lost and it becomes an immediate concrete experience” (Skårderud,

2007a). These observations have been interpreted as evidence for the impairment of the reflective function of the mind, namely “the psychological processes underlying the capacity to make mental representations”.

However, all these insights are not clear-cut and conclusive. Thus, the Linguistic profile of AN (and FED in general) remains, to date, mostly unexplored. Moreover, all the retrieved studies tackled verbal productions written in a language that belongs to the Germanic language group: English, German or Norwegian. Given the peculiar typological features of Italian language (e.g. at the morpho-syntactic level), these results cannot be readily generalized.

## 2. Materials and Methods

### 2.1 Rationale

Drawing on this wide body of clinical evidence and computational experiences, we hypothesize that (i) disturbances in self-perceived body image, black and white thinking and mood changes strongly associated with AN disorder can result in altered linguistic patterns; and (ii) these subtle modifications can be detected by means of NLP tools, acting as early proxy measures for the disorder. To test our hypothesis, the study will compare some short, written productions of AN patients with those of a control group, in order to identify possible distinctive linguistic features. To the best of our knowledge, this is the first work on linguistic profile of AN in Italian.

### 2.2 Data collection

The study was approved by the Ethics Committee of Azienda Ospedaliero-Universitaria di Bologna, Policlinico Sant’Orsola-Malpighi, Italy (prot. 683/2019/Oss/AOUBo). We enrolled 51 participants, ranging in age from 14 to 18: the sample is composed of an Anorexia Nervosa group (AN) and a Control Group (CG), with a ratio of 1:2. The AN group included 17 girls, recruited at the Regional Center of Eating disorders of the Child Neuropsychiatry Unit (Policlinico Sant’Orsola – Malpighi, University of Bologna) with a clinical diagnosis of Anorexia Nervosa according to national and international guidelines (American Psychiatric Association, 2013); 6 out of 17 show purging behavior, 12 have been experienced primary or secondary amenorrhea. The mean BMI of the group is 17.0. CG included 34 girls matched by gender, age and educational level (school grade/type of secondary school attended). Inclusion criteria are outlined in table 1, while table 2 summarizes the demographic characteristics of the sample.

AN	CG
- Age: 14-18	- Age: 14-18
- Diagnosis of Anorexia Nervosa (DSM-5)	- BMI $\geq$ 18.5
- fair level of communication skills in Italian (Language History Questionnaire)	- fair level of communication skills in Italian (Language History Questionnaire)
- written informed consent	- written informed consent

Table 1: Inclusion criteria for participant enrollment.

GROUP	N	AGE (mean $\pm$ sd)	YEARS OF EDUCATION (mean $\pm$ sd)
AN	17	16 $\pm$ 1.37	11.06 $\pm$ 1.34
CG	34	16 $\pm$ 1.35	11.15 $\pm$ 1.28

Table 2: Demographic characteristics of the sample.

Subjects were asked to produce three short written texts (around 10-15 lines long), in the presence of the experimenter:

1. personal task (-PER-): “How would you describe yourself? (Please, talk about your physical and personality traits, your hobbies etc.)”.
2. neutral task (-NEU-): “How do you usually spend time with your friends?”
3. description of a complex picture (-FIG-); the renowned black and white picture “Cookie theft” from the BDAE - Boston Diagnostic Aphasia Examination Battery (Goodglass et al., 2001) has been proposed as a stimulus (figure 1).

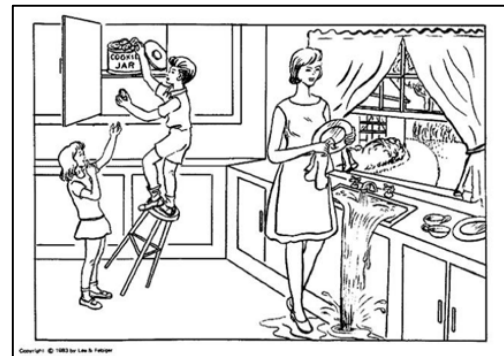


Figure 1: “The cookie theft” (Goodglass et al., 2001).

Language proficiency in Italian has been also assessed, by means of a short self-reported questionnaire. As a matter of fact, bilingualism and multilingualism are the norm rather than the exception in today’s Italy: this additional test aims at assessing both quality and quantity of bilingual experience, in order to remove from the sample poor productions due to scarce language exposure.

## 3. Data analysis

The handwritten texts have been converted into digital texts manually by the linguists. After the automatic tokenization of the transcripts, the corpus has been enriched by adding linguistic information at the lexical and morphosyntactic levels: all the sentences have been automatically PoS-tagged, lemmatized and syntactically parsed with the dependency model used by the Turin University Linguistic Environment – TULE<sup>2</sup> (Lesmo, 2007), based on the TUT - Turin University TreeBank tagset (Bosco et al., 2000). All the annotations have been manually checked by one linguist, in order to remove the errors introduced by the automatic tagging. The revision has been made by using the Dependency Grammar Annotator - DGA opensource

<sup>2</sup> <https://github.com/alexmazzei/TULE>

software<sup>3</sup> for an easy visualization and correction of TULE mistakes at any level (see figure 2).

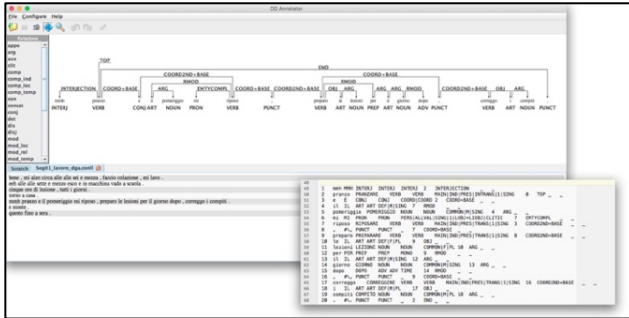


Figure 2: Dependency graph as shown by DGA and full utterance annotation in CoNLL-U format.

A multidimensional parameter analysis has been performed on the corpus: examining the relevant literature, we selected a wide range of linguistic/stylometric indexes to be tested in order to determine their relevance in the discrimination between AN subjects and normal weighted peers.

In addition, we used the software LIWC (Linguistic Inquiry and Word Count) (Chung and Pennebaker, 2007; Tausczik and Pennebaker, 2010; Agosti and Rellini, 2007), a text analysis program which counts the percentage of different lexical categories, in order to capture people’s social and psychological states (i.e. emotions, thinking styles, social concerns). The complete list for all the features considered in this study is reported in the Appendix A.

The Statistical significance of differences between AN and controls on all the indexes has been evaluated with the Kolmogorov–Smirnov non-parametric test, because of the small size of our corpus.

#### 4. Results

The focus of this study was the analysis of written texts of AN patients, in comparison to normal weighted peers. The study is still ongoing, with full results expected in 2021: therefore, the findings presented in this work are far from conclusive.

Age and schooling differences of the enrolled participants (table 2) are not statistically relevant at the Kolmogorov–Smirnov test; thus, the sample is well balanced on each variable.

Table 3 presents the number of words produced by the groups for each task. As corroborated by the statistical analysis, the three stimuli show different “elicitation power” (Kruskall-Wallis non-parametric test with Dunn’s multiple comparison,  $p$ -value  $< 0.001$ ): as a matter of fact, the “personal task” prompted richer responses in both samples.

Results for statistically relevant indexes are presented in table 4. For a complete picture of real values and a selection of our corpus, please refer to Appendix B, C and D.

task	AN (mean $\pm$ sd)	CG (mean $\pm$ sd)
task1 -PER-	98.63 $\pm$ 42.94	105.5 $\pm$ 35.05
task2 -NEU-	61.53 $\pm$ 40.98	68.56 $\pm$ 31.55
task 3 -FIG-	81.50 $\pm$ 40.02	77.15 $\pm$ 24.13
overall	80.22 $\pm$ 43.16	83.74 $\pm$ 34.18

Table 3: Text length, in tokens, produced on the three tasks by AN and CG subjects, shown as mean $\pm$ standard deviation.

FEATURES	task 1 -PER-	task 2 -NEU-	task 3 -FIG-	overall
LEX_ContDens			*	
LEX_PoS_ADV	*			
LEX_PoS_CONJ				*
LEX_PDEIXIS		*		
LEX_HonoreR			*	*
SYN_NPLENSD		*		
SYN_GRAPHDISTM				**
SYN_SLENM	*			**
SYN_SLENSD	*			*
LIWC_WPS		*		*
LIWC_SIXLTR			*	***
LIWC_DIC		*	***	
LIWC_PERCP		*		
LIWC_PRES			*	

Table 4: Results of the linguist analysis. The significant  $p$ -value is indicated for the corresponding feature and task, with \* $p < 0.05$ ; \*\* $p < 0.01$ ; \*\*\* $p < 0.001$ .

#### 5. Discussion and Concluding Remarks

Firstly, we notice that the most effective task is the description of a complex picture. This finding is not surprising: according to (Chung, 2007), linguistic tasks not directly pertaining to psychological and bodily states provide a non-reactive way to explore social and personality processes. However, aggregated tasks represent the best testing ground for the evaluation of subtle linguistic alteration: it seems trivial, but the simple merging of the three written texts allows to partially overcome the issue of data scarceness, increasing the sensitivity of the analysis.

From the qualitative point of view, syntactic reduction appears as the most relevant trait of AN productions. To this respect, several indexes emerged as statistically significant: sentence length mean and standard deviation, number of dependent elements linked to the noun, Global Dependency Distance and LIWC\_WPS, i.e. the number of tokens per sentence. Among the distinguishing lexical features of our cohort are: Content Density, i.e. the ratio of open-class words to closed-class words, Lexical Richness calculated as R – Honoré’s statistic, rate of Adverbs, Conjunctions and personal deixis, incidence of LIWC2007 Dictionary (LIWC\_DIC). At the semantic level, our data show lower incidence of lexical units related to perceptual processes (LIWC\_PERCP, i.e. multiple sensory and

<sup>3</sup> <http://medialab.di.unipi.it/Project/QA/Parser/DgAnnotator/>

perceptual dimensions associated with the five senses) in AN patients with respect to controls.

The most frequently described trait of AN, namely the abnormal use of first-person singular pronouns (Lyons et al., 2006; Wolf et al., 2013), is not confirmed by our data, as well as the plural ones, since the differences on LIWC\_1PS and LIWC\_1PP indexes are not statistically relevant. The analysis of temporal focus is controversial too: in contrast with the work of (Lyons et al., 2006), the written text of CG contains more present tense verbs (LIWC\_PRES), disconfirming the presumed attentional focus on the here and now. Furthermore, none of the readability features turn out to be statistically relevant, except for the usage of long (> 6 letter) words (LIWC\_SIXLTR).

However, these are preliminary data and additional evidences are needed to assess the actual reliability of linguistic parameters that have been proved to be probable proxy measures of AN. Moreover, due to the small size of the corpus, the order of the tasks was not counterbalanced across participants; this limitation should be tackled in the next administrations of the test.

Future works should also consider possible correlation between linguistic and clinical variables, such as diagnostic subtypes (“restricting” or “binge-eating/purging”), severity, physical signs and symptoms (e.g. amenorrhea), comorbidity (e.g. bipolar, depressive, anxiety, or obsessive-compulsive disorders), age of the onset and pharmacological treatment with Selective Serotonin Reuptake Inhibitors (e.g. fluoxetine, sertraline, fluvoxamine), anxiolytics (e.g. benzodiazepines) or antipsychotics (e.g. olanzapine, quetiapine).

If these preliminary results will be confirmed, the use of an automatic system that analyses and classifies patients' written productions can represent a promising approach for the identification of both overtly pathological and sub-clinical conditions.

## 6. Ethics Statement

All ethical principles of the Helsinki Declaration were followed.

The study was reviewed and approved by the Ethics Committee of Azienda Ospedaliero-Universitaria di Bologna, Policlinico Sant'Orsola-Malpighi, Italy (prot. 683/2019/Oss/AOUBo).

Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

## 7. Acknowledgements

The authors are deeply grateful to the participants and their families, who freely gave their time to participate in the study. They are also indebted to the principals of the secondary schools involved in the study: Elena Ugolini (liceo M.Malpighi, Bologna), Vincenzo Manganaro (IIS B. Scappi, Castel San Pietro Terme) and Giovanna Degli Esposti (liceo Manzoni, Bologna). The precious help of Chiara Gianollo and Annalisa Raffone is also acknowledged.

## 8. Author contribution

GM: literature review, linguistic data collection and annotation, statistical analysis; GG: literature review, methodology, statistical analysis, writing; VC: clinical data collection; FT: automatic feature extraction; EM, PG, FS, FM, VF: collaborators; AP: supervision and project administration.

## 9. Bibliographical References

- Adornetti, I. (2018). *Patologie del linguaggio e della comunicazione*. Carocci, Roma, Italy.
- Agosti, A. and Rellini, A. (2007). *The Italian LIWC dictionary*. Technical Report 1, Austin, TX.
- Al-Mosaiwi, M. and Johnstone, T. (2018). In an absolute state: elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. *Clinical Psychological Science*, 6(4):529–542.
- Altmann, L.J. and Troche, M. (2011). High-level language production in Parkinson's disease: A review. *Parkinson's Disease*, 238956.
- American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*. APA, Washington, DC.
- Arcelus, J., Mitchell, A.J., Wales, J., and Nielsen, S. (2011). Mortality rates in patients with Anorexia Nervosa and other Eating Disorders: A Meta-analysis of 36 Studies. *Archives of General Psychiatry*, 68(7):724–731.
- Arntz, A., Hawke, L., Bamelis, L., Spinhovend, P., and Molendijk, M. (2012). Changes in natural language use as an indicator of psychotherapeutic change in personality disorders. *Behaviour Research and Therapy*, 50:191–202.
- Asgari, M., Kaye, J., and Dodge, H. (2017). Predicting Mild Cognitive Impairment from spontaneous spoken utterances. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 3(2):219–228.
- Bambini, V., Arcara, G., Bechi, M., Buonocore, M., Cavallo, R., and Bosia, M. (2016). The communicative impairment as a core feature of Schizophrenia: frequency of pragmatic deficit, cognitive substrates, and relation with quality of life. *Comprehensive Psychiatry*, 71:106–120.
- Becker, A.E., Grinspoon, S.K., Klibanski, A., and Herzog, D.B. (1999). Eating disorders. *New England Journal of Medicine*, 340(14):1092–1098.
- Beltrami, D., Gagliardi, G., Rossini Favretti, R., Ghidoni, E., Tamburini, F., and Calzà, L. (2018). Speech analysis by Natural Language Processing techniques: A possible tool for very early detection of cognitive decline? *Frontiers in Aging Neuroscience*, 10:369.
- Bernard, J., Baddeley, J., Rodriguez, B., and Burke, P. (2016). Depression, language, and affect: an examination of the influence of baseline depression and affect induction on language. *Journal of Language and Social Psychology*, 35(3):317–326.
- Boschi, V., Catricalà, E., Consonni, M., Chesi, C., Moro, A., and Cappa, S.F. (2017). Connected Speech in Neurodegenerative Language Disorders: a review. *Frontiers in Psychology*, 8:1–21.
- Bosco, C., Lombardo, V., Vassallo, D., and Lesmo, L. (2000). Building a treebank for Italian: a data-driven annotation schema. In *Proceedings of the Second International Conference on Language Resources and*

- Evaluation (LREC-2000)*, pages 99-105, ELRA - European Language Resources Association, Paris, France.
- Brockmeyer, T., Holtforth, M.G., Bents, H., Herzog, W. and, Friederich, H. (2013). Lower body weight is associated with less negative emotions in sad autobiographical memories of patients with Anorexia Nervosa. *Psychiatry Research*, 210:548–552.
- Brockmeyer, T., Zimmermann, J., Kulesa, D., Hautzinger, M., Bents, H., Friederich, H., Herzog, W., and Backenstrass, M. (2015). Me, myself, and I: self-referent word use as an indicator of self-focused attention in relation to depression and anxiety. *Frontiers in Psychology*, 6:1564.
- Brunet, E. (1978). *Le Vocabulaire de Jean Giraudoux. Structure et évolution*. Slatkine, Geneva, Switzerland.
- Bucks, R.S., Singh, S., Cuerden, J.M., and Wilcock, G.K. (2000). Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance. *Aphasiology*, 14(1):71–91.
- Cantos-Gómez, P. (2009). Featuring linguistic decline in Alzheimer’s disease: a corpus-based approach. In M. Mahlberg, et al., editors, *Proceedings of the Corpus Linguistics Conference 2009 (CL2009)*, article 218, University of Liverpool, Liverpool, UK.
- Catricalà, E., Boschi, V., Cuoco, S., Galiano, F., Picillo, M., Gobbi, E., Miozzo, A., Chesi, C., Esposito, V., Santangelo, G., Pellicchia, M.T., Borsa, V.M., Barone, P., Garrard, P., Iannaccone, S., and Cappa, S.F. (2019). The language profile of Progressive Supranuclear Palsy. *Cortex*, 115:294–308.
- Chung, C., and Pennebaker, J. (2007). The psychological functions of function words. In K. Fiedler, editor, *Social Communication*, pages 343–359. Psychology Press, New York.
- Cole, T., Flegal, K., Nicholls, D., and Jackson, A. (2007). Body mass index cut offs to define thinness in children and adolescents: international survey. *BMJ: British Medical Journal*, 335(7612):194.
- Crystal, D. (1981). *Clinical Linguistics*. Springer-Verlag, Wien, Austria.
- De Mauro, T. (2000). *Il dizionario della lingua italiana*. Paravia, Torino, Italy.
- Dovetto, F.M. (2015). *Uso delle parole nella schizofrenia*. In L. Mariottini, editor, *Identità e discorsi. Studi offerti a Franca Orletti*, pages 223–236. RomaTre Press, Roma, Italy.
- Edwards, T., and Holtzman, N. (2017). A meta-analysis of correlations between depression and first person singular pronoun use. *Journal of Research in Personality*, 68:63–68.
- Enckell, H. (2002). *Metaphor and the psychodynamic functions of the mind*. Ph.D. thesis, Kuopion Yliopisto, Kuopio, Finland.
- Espeset, E., Gulliksen, K., Nordbø, R.H., Skårderud, F., and Holte, A. (2012). Fluctuations of body images in anorexia nervosa: Patients’ perception of contextual triggers. *Clinical Psychology and Psychotherapy*, 19(6):518–530.
- Fraser, K.C., Lundholm Fors, K., Eckerström, M., Öhman F., and Kokkinakis, D. (2019a). Predicting MCI status from multimodal language data using Cascaded Classifiers. *Frontiers in Aging Neuroscience*, 11:205.
- Fraser, K.C., Lundholm Fors, K., and Kokkinakis, D. (2019b). Multilingual word embeddings for the assessment of narrative speech in Mild Cognitive Impairment. *Computer Speech & Language*, 53:121–139.
- Gagliardi, G. (2014). *Validazione dell’ontologia dell’azione IMAGACT per lo studio e la diagnosi del Mild Cognitive Impairment*. PhD Thesis, Università degli Studi di Firenze.
- Gagliardi, G. (2019). *Linguistica per le professioni sanitarie*. Patron, Bologna, Italy.
- Gagnon, M., Barrette, J., and Macoir, J. (2018). Language disorders in Huntington disease: A systematic literature review. *Cognitive and behavioral neurology: official journal of the Society for Behavioral and Cognitive Neurology*, 4(31):179–192.
- Gavin, J., Rodham, K., and Poyer, H. (2008). The presentation of “pro-anorexia” in online group interactions. *Qualitative Health Research*, 18:325–333.
- Goodglass, H., Kaplan, E., and Barresi, B. (2001). *The Boston Diagnostic Aphasia Examination (BDAE)*. Lippincott Williams & Wilkins.
- Gosztolya, G., Vincze, V., Tóth, L., Pákási, M., Kálmán, J., and Hoffmann, I. (2019). Identifying Mild Cognitive Impairment and Mild Alzheimer’s Disease based on spontaneous speech using ASR and linguistic features. *Computer Speech & Language*, 53:181–197.
- Harris, C., and Barraclough, B. (1998). Excess mortality of mental disorder. *British Journal of Psychiatry*, 173(1):11–53.
- Herzog, D.B., Nussbaum, K.M., and Marmor, A.K. (1996). Comorbidity and outcome in eating disorders. *Psychiatric Clinics of North America*, 19(4):843–859.
- Hoek, H.W. (1993). Review of the epidemiological studies of Eating Disorders. *International Review of Psychiatry*, 5(1):61–74.
- Holmes, D.I. and Singh, S. (1996). A stylometric analysis of conversational speech of aphasic patients. *Literary and Linguistic Computing*, 11(3):133–140.
- Honoré, A. (1979). Some Simple Measures of Richness of Vocabulary. *Association of Literary and Linguistic Computing Bulletin*, 7:172–177.
- Lesmo, L. (2007). Il parser basato su regole del Gruppo NLP dell’Università di Torino. *Intelligenza Artificiale*, IV(2):46–47.
- Lyons, E., Mehl, M., and Pennebaker, J. (2006). Pro-anorexics and recovering anorexics differ in their linguistic internet self-presentation. *Journal of Psychosomatic Research*, 60:253–256.
- March, E.G., Wales, R., and Pattison, P. (2006). The uses of nouns and deixis in discourse production in Alzheimer’s disease. *Journal of Neurolinguistics*, 19:311–340.
- Marini, A., and Carlomagno, S. (2004). *Analisi del discorso e patologia del linguaggio*. Springer-Verlag, Milano, Italy.
- Montemurro, S., Mondini, S., Signorini, M., Marchetto, A., Bambini, V., and Arcara, G. (2019). Pragmatic language disorder in Parkinson’s disease and the potential effect of cognitive reserve. *Frontiers in Psychology*, 10:1220.
- Ramirez-Esparza, N., Chung, C., Kacwicz, E., and Pennebaker, J. (2008). The psychology of word use in depression forums in English and in Spanish: Testing two text analytic approaches. In E. Adar, et al., editors, *Proceedings of the Second International Conference on*

- We-blogs and Social Media, ICWSM 2008*, pages 102–110. AAAI Press, Menlo Park, CA.
- Roark, B., Mitchell, M., and Hollingshead, K. (2007). Syntactic complexity measures for detecting Mild Cognitive Impairment. In K. Bretonnel Cohen, et al., editors, *Proceedings of the Workshop BioNLP 2007: Bio-logical, translational, and clinical language processing*, pages 1–8, ACL - Association for Computational Linguistics.
- Roark, B., Mitchell, M., Hosom, J.P., Hollingshead, K., and Kaye, J.A. (2011). Spoken Language Derived Measures for Detecting Mild Cognitive Impairment. *IEEE Transactions on Audio Speech, and Language Processing*, 19(7):2081–2090.
- Rubin, D., and Greene, K. (1992). Gender-typical style in written language. *Research in the Teaching of English*, 26:7–40.
- Skårderud, F. (2007a). Eating one’s words, part I: ‘concretised metaphors’ and reflective function in Anorexia Nervosa – an interview study. *European Eating Disorders Review*, (3):163–74.
- Skårderud, F. (2007b). Eating one’s words, part II: The embodied mind and reflective function in Anorexia Nervosa – theory. *European Eating Disorders Review*, 15(4): 243–252.
- Smink, F.R.E., van Hoeken, D., and Hoek, H.W. (2012). Epidemiology of Eating Disorders: Incidence, prevalence and mortality rates. *Current Psychiatry Reports*, 14(4):406–414.
- Smirnova, D., Cumming, P., Sloeva, E., Kuvshinova, N., Romanov, D., and Nosachev, G. (2018). Language patterns discriminate mild depression from normal sadness and euthymic state. *Frontiers in Psychiatry*, 9:105.
- Snowdon, D.A., Kemper, S.J., Mortimer, J.A., Greiner, L.H., Wekstein, D.R., and Markesbery, W.R. (1996). Linguistic Ability in Early Life and Cognitive Function and Alzheimer’s Disease in Late Life: Findings from the Nun Study. *The Journal of the American Medical Association*, 275:528–532.
- Spinczyk, D., Nabrdalik, K., and Rojewska, K. (2018). Computer aided sentiment analysis of anorexia nervosa patients’ vocabulary. *Biomedical engineering online*, 17(1):19.
- Szmrecsányi, B.M. (2004). On Operationalizing Syntactic Complexity. In G. Purnelle, et al., editors, *Proceedings of the 7th International Conference on Textual Data Statistical Analysis*, pages 1031–1038, Presses Universitaires de Louvain, Louvain-la-Neuve, Belgium.
- Tausczik, Y., and Pennebaker, J. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.
- Themistocleous, C., Eckerström, M., and Kokkinakis, D. (2018). Identification of Mild Cognitive Impairment from speech in Swedish using deep sequential neural networks. *Frontiers in Neurology*, 9:975.
- Tóth, L., Hoffmann, I., Gosztolya, G., Vincze, V., Sztatlóczi, G., Bánreti, Z., Pákáski, M., and Kálmán, J. (2018). A Speech Recognition-based Solution for the Automatic Detection of Mild Cognitive Impairment from Spontaneous Speech. *Current Alzheimer Research*, 15:1–10.
- Trotzek, M., Koitka, S., and Friedrich, C.M. (2018). Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *IEEE Transactions on Knowledge and Data Engineering*, 32(3): 588–601.
- Vigorelli, P. (2004). *La conversazione possibile con il malato Alzheimer*. Franco Angeli, Milano, Italy.
- Vincze, V., Gosztolya, G., Tóth, L., Hoffmann, I., Sztatlóczi, G., Bánreti, Z., Pákáski, M., and Kálmán, J. (2016). Detecting Mild Cognitive Impairment by Exploiting Linguistic Information from Transcripts. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 181–187, ACL - Association for Computational Linguistics.
- Wolf, M., Theis, F., and Kordy, H. (2013). Language use in Eating Disorder blogs: psychological implications of social online activity. *Journal of Language and Social Psychology*, 32(2):212–226.
- World Health Organization. (1993). *The ICD-10 classification of mental and behavioural disorders*. WHO, Geneva, Switzerland.
- World Health Organization. (1995). *Physical status: the use and interpretation of anthropometry*. WHO, Geneva, Switzerland.
- Zimmermann, J., Brockmeyer, T., Hunn, M., Schauenburg, H., and Wolf, M. (2017). First-person pronoun use in spoken language as a predictor of future depressive symptoms: preliminary evidence from a clinical sample of depressed patients. *Clinical Psychology and Psychotherapy*, 24:384–391.

APPENDIX A: LIST OF FEATURES TAKEN INTO CONSIDERATION IN THIS STUDY		
	INDEX	DESCRIPTION & BIBLIOGRAFIC REFERENCES
Lexical features	Text length LEX_NW	Number of tokens
	Content Density LEX_ContDens	The ratio of open-class tokens to closed-class tokens (Roark et al., 2011).
	Part-of-Speech rate LEX_PoS_*	The average rate of occurrence for each Part-of-Speech (PoS) category (Holmes and Singh, 1996; Bucks et al., 2000).
	Reference Rate to Reality LEX_RefRRReal	The ratio of the total number of nouns to the total number of verbs (Vigorelli, 2004).
	Personal, Spatial and Temporal Deixis rate LEX_*DEIXIS	The rate of deictic expressions in the written text w.r.t. the total number of tokens (March et al., 2006; Cantos-Gómez, 2009).
	Relative pronouns and negative adverbs rate LEX_RPRO	The rate of relative pronouns.
	Lexical Richness LEX_TTR; LEX_BrunetW; LEX_HonoreR	This class of measures quantifies the richness of vocabulary/lexical diversity (Holmes and Singh, 1996; Brunet, 1978; Honoré, 1979): - <i>TTR, Type-Tokens Ratio</i> - <i>W, Brunet's Index</i> - <i>R, Honore's Statistic</i>
	Action Verbs rate LEX_ACTVRB	The metric probes the rate of action verbs (i.e. verbs referring to physical action, like "to put", "to run", "to eat") in the texts. (Gagliardi, 2014).
	Frequency-of-use LEX_DM_F	Mean frequency-of-use weight among words extracted from the De Mauro's frequency list (De Mauro, 2000).
	Propositional Idea Density LEX_IDEAD	Idea density is the number of expressed propositions (i.e. distinct facts or notions contained in a text) divided by the number of tokens (Snowdon et al., 1996; Roark et al., 2011).
Syntactic features	Number of dependent elements linked to the noun SYN_NPLENM; SYN_NPLENSD	The feature explores Noun Phrase complexity, counting the number of dependent elements linked to the head (e.g. Adjectives, Relative clauses...). Mean and Std. Deviation were taken into account.
	Global Dependency Distance SYN_GRAPHDISTM; SYN_GRAPHDISTSD	Given the memory overhead of long distance dependencies, the feature quantifies the difficulty in syntactic processing (Roark et al., 2007; Roark et al., 2011). Mean and Std. Deviation were taken into account.
	Syntactic complexity SYN_ISynCompI	Syntactic complexity is established by counting the linguistic tokens that can be considered to telltale signs of increased grammatical subordinateness and embeddedness, such as subordinating conjunctions, WH- pronouns, verb forms, both finite and non-finite and noun phrases. (Szmrecsányi, 2004).
	Syntactic embeddedness SYN_MAXDEPTHM; SYN_MAXDEPTHSD	The maximum "depth" of the dependency structure. Mean and Std. Deviation were taken into account.
	Sentence length SYN_SLENM; SYN_SLENSD	The average number of tokens for sentence. Mean and Std. Deviation were taken into account.
LIWC	Linguistic processes	Total words count (WC), Words per sentence (WPS), Words > 6 letters (SIXLTR), Dictionary words count (DIC)
	Function Words	1 <sup>st</sup> person singular (1PS), 1 <sup>st</sup> person plural (1PP), 2 <sup>nd</sup> person singular (2PS), 2 <sup>nd</sup> person plural (2PP), 3 <sup>rd</sup> person singular (3PS), 3 <sup>rd</sup> person plural (3PP), Negations (NEG), Past tense (PST), Present tense (PRES), Future tense (FUT), Gerund (GER), Conditional mood (COND), Passive voice (PASS), Past Participle (PP), Transitivity (TRAN)
	Affective processes (AFFP)	Positive emotions (+EMO), Negative emotions (-EMO), Anxiety (ANX), Anger (ANG), Sadness (SAD)
	Cognitive Processes (COGP)	Insight (INS), Cause (CAU), Discrepancies (DISCR), Tentativeness (TENT), Certainty (CERT), Inhibition (INH), Inclusive (INCL), Exclusive (EXCL)
	Perceptual processes (PERCP)	See (SEE), Hear (HEAR), Feel (FEEL)
	Biological processes (BIOP)	Body (BODY), Health (HLT), Ingestion (ING)
	Personal concerns (PERSC)	Work (WORK), School (SCHOOL), Death (DEATH), Achievement (ACH), Leisure (LEIS), Home (HOME), Sport (SPORT)
Psychological processes (PSYP)	Family (FAM), Friends (FR), Humans (HUM), Social processes (SOC)	

APPENDIX B: RESULTS OF LEXICAL AND SYNTACTIC FEATURES EXTRACTION (mean ± standard deviation)								
Feature	task 1 -PER-		task 2 -NEU-		task 3 -FIG-		overall	
	AN	CG	AN	CG	AN	CG	AN	CG
LEX_NW	98.63 ± 42.94	105.5 ± 35.05	61.53 ± 40.98	68.56 ± 31.55	81.50 ± 40.02	77.15 ± 24.13	80.22 ± 43.16	83.74 ± 34.18
LEX_ContDens	1.32±0.19	1.37±0.17	1.20± 0.23	1.14±0.19	1.17±0.15	1.07±0.14	1.22±0.20	1.19±0.21
LEX_PoS_*								
ADJ	0.13±0.02	0.11±0.03	0.06±0.03	0.05±0.03	0.06±0.03	0.04±0.03	0.08±0.04	0.07±0.04
ADV	0.10±0.04	0.12±0.04	0.09±0.05	0.11±0.05	0.07±0.04	0.06±0.04	0.09±0.05	0.10±0.05
ART	0.06±0.03	0.07±0.02	0.06±0.03	0.05±0.02	0.11±0.03	0.12±0.02	0.08±0.04	0.08±0.04
CONJ	0.08±0.03	0.09±0.02	0.08±0.04	0.09±0.03	0.06±0.03	0.07±0.02	0.07±0.03	0.08±0.03
DATE	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.01	0.00±0.00	0.00±0.01	0.00±0.00
INTERJ	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
NOUN	0.14±0.03	0.13±0.03	0.18±0.06	0.17±0.04	0.19±0.03	0.20±0.03	0.17±0.05	0.17±0.04
NUM	0.01±0.01	0.01±0.01	0.00±0.01	0.00±0.01	0.01±0.01	0.02±0.01	0.01±0.01	0.01±0.01
PHRAS	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
PREDET	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.01	0.00±0.00	0.00±0.00
PREP	0.08±0.03	0.07±0.03	0.13±0.06	0.14±0.04	0.11±0.04	0.11±0.03	0.12±0.05	0.11±0.05
PRON	0.07±0.03	0.07±0.02	0.06±0.05	0.07±0.04	0.07±0.04	0.08±0.03	0.07±0.04	0.07±0.03
VERB	0.20±0.04	0.21±0.03	0.20±0.06	0.20±0.03	0.21±0.04	0.22±0.04	0.20±0.05	0.21±0.03
LEX_RefRRReal	0.72±0.23	0.67±0.17	1.04±0.71	0.92±0.33	0.99±0.27	0.96±0.24	0.93±0.47	0.85±0.28
LEX_PDEIXIS	0.04±0.03	0.04±0.02	0.03±0.03	0.04±0.03	0.03±0.01	0.03±0.02	0.03±0.02	0.04±0.02
LEX_SDEIXIS	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.01±0.01	0.00±0.00	0.00±0.01
LEX_TDEIXIS	0.01±0.01	0.01±0.01	0.01±0.01	0.01±0.01	0.00±0.01	0.00±0.00	0.01±0.01	0.00±0.01
LEX_RPRO	0.01±0.01	0.01±0.01	0.01±0.01	0.01±0.01	0.02±0.02	0.02±0.02	0.01±0.02	0.01±0.02
LEX_NEGADV	0.02±0.01	0.02±0.01	0.00±0.01	0.01±0.01	0.01±0.02	0.01±0.01	0.01±0.01	0.01±0.01
LEX_TTR	0.69±0.06	0.69±0.06	0.79±0.07	0.75±0.08	0.76±0.07	0.73±0.07	0.75±0.08	0.73±0.07
LEX_BrunetW	9.63±1.01	9.85±0.68	8.38±1.23	8.90±0.98	9.15±0.78	9.25±0.76	9.04±1.12	9.33±0.90
LEX_HonoreR	2408.7±659.1	2197.3± 480.0	2325.4±679.8	2172.6±703.0	2326.0±765.8	2026.1±747.5	2351.8±692.2	2131.9±652.0
LEX_ACTVRB	0.04±0.03	0.03±0.02	0.05±0.02	0.05±0.03	0.07±0.03	0.52±0.03	0.05±0.03	0.05±0.03
LEX_DM_F	3.08±0.57	3.11±0.55	2.12±0.76	2.10±0.67	1.54±0.98	1.24±0.55	2.22±1.01	2.15±0.96
LEX_IDEAD	0.59±0.05	0.61±0.04	0.56±0.07	0.59±0.06	0.54±0.05	0.52±0.04	0.56±0.06	0.57±0.06
SYN_NPLENM	2.39±0.70	2.46±0.58	1.76±0.68	1.97±0.59	2.10±0.59	1.69±0.49	2.08±0.69	2.04±0.63
SYN_NPLENSD	1.95±0.80	1.84±0.61	1.23±1.05	1.49±0.68	1.98±0.65	1.51±0.91	1.72±0.90	1.61±0.76
SYN_GRAPHDISTM	1.34±0.29	1.46±0.20	1.59±0.35	1.67±0.21	1.66±0.43	1.72±0.29	1.54±0.38	1.62±0.26
SYN_GRAPHDISTD	0.37±0.15	0.44±0.16	0.26±0.17	0.34±0.24	0.45±0.25	0.46±0.32	0.36±0.21	0.42±0.25
SYN_ISynCompl	0.34±0.03	0.36±0.04	0.41±0.05	0.42±0.05	0.39±0.04	0.41±0.06	0.38±0.05	0.40±0.06
SYN_MAXDEPTHM	7.60±2.75	7.40±1.37	7.16±1.53	8.50±3.29	8.73±2.64	8.87±3.86	7.85±2.42	8.25±3.07
SYN_MAXDEPTHD	2.48±1.44	2.52±1.16	1.87±1.42	2.07±1.90	2.56±1.55	3.04±1.73	2.30±1.48	2.54±1.66
SYN_SLENM	17.44±6.92	19.27±4.93	19.54±5.30	25.84±11.81	24.50±10.26	26.38±11.05	20.63±8.26	23.83±10.20
SYN_SLENSD	5.64±2.89	8.73±4.59	5.58±4.60	7.60±7.25	8.37±5.26	9.41±6.42	6.58±4.52	8.58±6.17

APPENDIX C: RESULTS OF LIWC FEATURES EXTRACTION (mean ± standard deviation)								
Feature	task 1 -PER-		task 2 -NEU-		task 3 -FIG-		overall	
	AN	CG	AN	CG	AN	CG	AN	CG
WC	85.76±36.64	92.76±31.20	53.71±35.89	61.67±29.40	72.41±35.17	70.03±21.36	70.63±37.60	74.82±30.40
WPS	15.68±6.55	16.77±3.99	16.67±5.04	23.17±11.39	24.48±9.51	24.64±11.43	18.94±8.16	21.52±10.10
SIXLTR	26.23±5.31	22.91±3.94	27.71±7.29	24.38±6.10	28.18±6.60	25.43±4.30	27.37±6.38	24.24±4.93
DIC	63.09±5.01	65.90±4.95	59.77±5.84	64.20±7.66	59.57±4.54	67.16±5.54	60.81±5.31	65.76±6.22
1PS	12.74±2.98	14.10±2.94	5.09±3.86	6.21±5.10	1.87±3.22	1.53±1.25	6.57±5.67	7.28±6.25
1PP	0.00±0.00	0.12±0.35	6.40±5.25	5.67±3.92	0.36±0.73	0.10±1.14	2.25±4.22	2.26±3.39
2PS	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
2PP	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
3PS	0.04±0.16	0.02±0.13	0.00±0.00	0.00±0.00	0.16±0.45	0.19±0.47	0.06±0.28	0.07±0.29
3PP	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
NEG	2.44±2.27	2.29±1.33	0.26±0.74	1.23±1.80	1.15±1.52	1.49±1.41	1.28±1.84	1.67±1.58
PST	0.25±0.74	0.08±0.35	0.58±1.34	0.07±0.43	0.16±0.54	0.24±0.73	0.33±0.94	0.13±0.53
PRES	13.46±3.80	15.08±2.89	10.33±5.58	11.19±4.89	7.08±3.18	9.54±2.49	10.29±4.98	11.94±4.24
FUT	0.00±0.00	0.04±0.20	0.00±0.00	0.00±0.00	0.14±0.40	0.34±0.63	0.05±0.24	0.12±0.41
GER	0.05±0.20	0.03±0.20	0.35±0.71	0.19±0.75	1.70±2.41	1.98±2.14	0.67±1.60	0.73±1.57
COND	0.78±1.08	0.31±0.55	0.07±0.29	0.05±0.21	0.28±0.69	0.4±0.72	0.38±0.80	0.25±0.55
PASS	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
PP	0.56±0.90	0.4±0.58	0.50±0.94	0.47±1.01	0.95±1.14	0.96±1.06	0.67±1.00	0.61±0.93
TRAN	0.18±0.40	0.35±0.70	0.99±1.50	0.20±0.73	1.55±1.27	2.03±1.17	0.91±1.27	0.86±1.21
AFFP	8.06±3.06	7.92±3.16	2.29±2.40	3.52±3.01	1.75±1.98	1.2±2.12	4.03±3.80	4.21±3.94
+EMO	4.60±2.72	3.86±2.28	1.02±1.47	0.97±1.47	0.48±0.80	0.12±0.40	2.03±2.58	1.65±2.25
-EMO	1.67±1.27	1.71±1.51	0.46±1.35	0.81±1.34	0.97±1.29	0.81±1.77	1.03±1.37	1.11±1.59
ANX	0.45±0.84	0.26±0.52	0.00±0.00	0.07±0.24	0.06±0.26	0.10±0.33	0.17±0.54	0.14±0.39
ANG	0.70±1.03	0.77±1.24	0.00±0.00	0.06±0.37	0.32±0.72	0.19±0.69	0.34±0.77	0.34±0.89
SAD	0.12±0.36	0.42±0.71	0.20±0.66	0.41±0.95	0.26±0.62	0.43±0.91	0.20±0.56	0.42±0.85
COGP	4.91±2.59	4.92±2.95	1.32±2.18	2.81±2.72	3.04±2.10	4.44±2.76	3.09±2.70	4.06±2.93
INS	1.37±1.28	1.62±1.70	0.55±1.38	0.84±1.39	1.05±1.15	2.24±2.18	0.99±1.29	1.56±1.86
CAU	0.45±0.67	0.38±0.75	0.00±0.00	0.38±0.77	0.33±0.63	0.34±0.64	0.26±0.55	0.37±0.71
DISCR	1.77±1.57	2.14±1.65	0.35±0.76	0.99±1.54	0.08±0.95	1.10±1.21	0.97±1.27	1.41±1.55
TENT	3.08±1.90	3.55±2.13	3.21±2.05	4.33±2.65	1.75±1.69	1.83±1.79	2.68±1.96	3.24±2.44
CERT	0.97±1.12	1.49±1.54	0.51±1.11	0.88±1.36	0.46±1.13	0.38±0.68	0.65±1.12	0.91±1.32
INH	0.44±0.69	0.30±0.48	0.10±0.41	0.08±0.36	0.03±0.13	0.15±0.44	0.19±0.49	0.18±0.44
INCL	0.78±1.06	1.07±1.17	1.20±1.48	0.78±1.28	0.50±0.84	1.02±1.17	0.83±1.17	0.96±1.20
EXCL	3.83±2.17	4.80±2.40	4.17±3.00	5.32±3.56	3.70±2.34	4.88±2.17	3.90±2.49	5.00±2.76
PERCP	3.12±2.04	2.71±1.52	1.55±1.95	2.80±2.09	0.85±0.90	1.39±1.48	1.84±1.93	2.30±1.82
SEE	1.79±1.51	1.14±0.96	0.63±1.74	0.55±0.95	0.52±0.78	0.77±1.00	0.98±1.49	0.82±0.99
HEAR	0.73±1.18	1.13±1.18	0.74±1.10	1.83±1.61	0.00±0.00	0.00±0.00	0.49±0.98	0.99±1.37
FEEL	0.14±0.31	0.20±0.62	0.00±0.00	0.09±0.39	0.09±0.28	0.27±0.55	0.08±0.24	0.19±0.53
BODY	4.32±1.87	3.98±2.38	0.30±0.70	1.15±1.57	1.88±1.27	2.55±1.51	2.17±2.14	2.56±2.18
HLT	0.76±0.68	0.52±0.56	0.00±0.00	0.04±0.26	0.03±0.13	0.19±0.57	0.26±0.52	0.25±0.52
ING	0.79±1.31	0.36±0.91	0.47±0.91	0.61±1.27	2.12±1.46	2.46±1.08	1.13±1.42	1.15±1.44
WORK	0.13±0.37	0.15±0.36	0.30±0.96	0.04±0.21	0.00±0.00	0.00±0.00	0.14±0.60	0.06±0.25
SCHOOL	0.36±0.76	0.32±1.04	0.85±2.05	0.23±0.64	0.00±0.00	0.00±0.00	0.40±1.28	0.18±0.71
DEATH	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
ACH	0.38±0.79	0.43±0.79	0.13±0.54	0.22±0.57	0.39±0.76	0.45±0.70	0.30±0.70	0.37±0.69
LEIS	1.50±1.28	0.96±1.49	2.36±1.82	3.10±1.96	1.60±1.93	1.18±1.00	1.82±1.72	1.75±1.80
HOME	0.70±1.01	0.30±0.62	0.77±1.17	0.89±1.01	1.52±1.9	1.18±1.00	1.00±1.44	0.79±0.96
SPORT	0.00±0.00	0.10±0.40	0.00±0.00	0.35±1.23	0.00±0.00	0.00±0.00	0.00±0.00	0.15±0.76
FAM	0.33±0.54	0.17±0.41	0.00±0.00	0.06±0.32	3.30±2.08	3.50±1.88	1.21±1.93	1.24±1.95
FR	0.80±1.10	0.77±0.86	2.10±2.10	2.01±1.80	0.03±0.13	0.00±0.00	0.98±1.60	0.92±1.41
HUM	2.08±1.60	2.07±1.37	0.77±1.75	0.18±0.55	3.64±2.18	3.88±2.34	2.17±2.18	2.04±2.19
SOC	4.95±2.90	4.75±2.33	8.69±5.41	8.60±5.09	8.09±2.32	9.01±2.63	7.24±4.06	7.45±4.03

## APPENDIX D: EXAMPLES FROM THE CORPUS

### Task 1 -PER-

#### AN, 18 years old

Sono una ragazza alta, capelli lunghi, occhi verdi e lentiggini. Sono simpatica, irascibile solare ma a volte cupa, solitaria e timida. Tante volte sono molto testarda e sfacciata, ma lo riconosco. A volte sono molto orgogliosa. Mi piace stare con gli amici, il fidanzato, andare in discoteca, ma prevalentemente disegnare e cucinare. Adoro vedere le persone felici e soddisfatte del pasto che ho preparato. In compenso odio pulire, fare i compiti, ma con la musica migliora un po' la situazione.

*English transl.:* I'm a tall girl, with long hair, green eyes, and freckles. I'm funny, quick-tempered but with a sunny disposition, loner and shy. I'm often stubborn and cheeky, but I admit it. Sometimes I have too much pride. I like to stay with friends, my boyfriend, going to the disco, but above all drawing and cooking. I love seeing people happy and satisfied with what I cooked for them. At the same time, I hate cleaning, doing homework, but if I listen to music it gets better.

### Task 2 -NEU-

#### AN, 15 years old

Solitamente parliamo, spetteghiamo di alcune persone, e parliamo della scuola e dei professori. Quando usciamo andiamo in centro oppure ci incontriamo per fare i compiti.

*English transl.:* We usually talk, gossip about people, and chat about school and professors. When we go out, we meet downtown or to do homework.

### Task 3 -FIG-

#### AN, 15 years old

La prima cosa che ho pensato nel vedere l'immagine qui sopra, è come potesse quella donna apparire noncurante, quasi sorridente, della situazione caotica che la circonda. Ella stessa non si preoccupa del lavabo ormai pieno, da cui fuoriesce, a bagnare il pavimento da cucina, un'imponente mole d'acqua; anzi continua imperterrita strofinando un piatto, senza nemmeno scorgere il figlioletto che è prossimo a cadere dallo sgabello. Poco distanti, i bambini sono intenti rubare dalla dispensa dei biscotti, ma il maschietto rischia di cadere all'indietro; la bambina pare interessata solo ad afferrare il dolce che il fratello le porge con aria incerta, senza capire il pericolo che il compagno sta correndo. Questi due ladruncoli di cibi mi ricordano tanto le mie malsane abitudini di ingozzarmi di nascosto, ignorando qualsiasi circostanza, come fa la piccola nel disegno, e dimenticandomi di esistere all'infuori del semplice atto d'inghiottire e deglutire.

*English transl.:* The first thing I thought when I saw the picture up here was how this woman could be so careless as if she was making fun of the chaotic situation surrounding her. She doesn't care about the sink now full, from which an impressive amount of water pulls out pouring the floor of the kitchen; indeed, she insists on rubbing the dishes, without even noticing her little boy about to fall off the stool. Not far away, children are stealing biscuits from the pantry, but the little boy risks falling backward; the girl seems only interested in grasping the sweet her brother is offering her with uncertain air, without figuring out the risk her mate is running. These two little food thieves remind me so much of my unhealthy habits of gorging myself secretly, by ignoring any circumstances, as the little girl does in the drawing, and forgetting to exist apart from the simple fact of swallowing and swallowing.



# What Difference Does it Make? Early Dementia Detection Using the Semantic and Phonemic Verbal Fluency Task

Hali Lindsay<sup>1</sup>, Johannes Tröger<sup>1</sup>, Jan Alexandersson<sup>1</sup>, Alexandra König<sup>2</sup>

German Research Center for Artificial Intelligence (DFKI) Stuhlsatzenhausweg 3, Saarbrücken, Germany, 66125<sup>1</sup>

Memory Clinic, Association IA, CoBtek Lab CHU Université Côte d’Azur, France<sup>2</sup>

Hali.Lindsay@dfki.de

## Abstract

Verbal Fluency (VF) tasks are common cognitive tests that are used in the diagnosis of early stages of Dementia. There are two main types of VF tasks; Semantic Verbal Fluency (SVF) and Phonemic Verbal Fluency (PVF). While much work has been done on automatic diagnostic relevance of the SVF, research on the automatic analysis of the PVF task or a combination of both remains minimal. This paper explores methods of extracting features from the SVF and the PVF task according to clinical and temporal methods, as well as how combined within-subject features from both tasks can increment classification performance. We investigate an early diagnostic scenario with a binary classification between healthy controls (N=8) and those with mild cognitive impairment (N=19), a likely precursor to dementia. Synthetic data augmentation (SMOTE) is used to balance the data set and multiple machine learning models—logistic regression, support vector machines with linear and radial basis function, and a multi-layer perceptron—are used to evaluate the features. The best performance comes from combining SVF, PVF and novel joint within-subject features (AUC > 0.90) for multiple machine learning methods.

**Keywords:** Phonemic Verbal Fluency, Semantic Verbal Fluency, Machine Classification, Clustering, mild cognitive impairment

## 1. Introduction

Verbal fluency (VF) tasks are amongst the most widely applied neuropsychological tests for the assessment of neurocognitive disorders. They are especially used for the diagnosis of different stages of dementia, ranging from very mild or even prodromal forms to clinical forms like in Alzheimer’s disease. The main strength of VF tasks are their ease of use (no testing material required and fully speech-based interaction) and brevity (1-2 minutes) given a high sensitivity for above-mentioned diagnostic purposes. Despite their traditional wide adoption in clinical and diagnostic practice, there is an ongoing scientific discussion regarding what verbal fluency tasks actually measure in terms of neurocognitive functions. However, multiple studies show that VF tasks generate rich variance stemming from the interplay of multiple neurocognitive functions including executive functions (EF) as well as memory and language components. Differentiating between them and identifying those VF contributors is crucial to understand how VF could be used to differentiate between multiple dementia sub-forms. VF as a test category comprises two major versions, the semantic verbal fluency (SVF) and the phonemic verbal fluency (PVF). Both follow similar rules: One has to produce as much different words as possible within a given timeframe and a given constraint. In the SVF the constraint is that all produced words should belong to one semantic category (e.g. animals) and within the PVF the constraint is that all produced words should start with one letter (e.g. S).

Methodologically, it is best clinical practice to test for both VF: SVF and PVF. In this context, multiple studies report a Verbal Fluency Discrepancy, meaning a performance advantage in the SVF compared to the PVF. This often is explained by the fact that in SVF one can follow associations for word production, whereas in PVF associations additionally have to be monitored for their phonemic fit

which puts additional EF demands on the testee. Therefore, in general, performance of healthy elderlies is better in the SVF than in PVF and this effect is preserved over aging (Vaughan et al., 2016). Patients with Mild Cognitive Impairment (MCI) of the amnesic type (precursor of Alzheimer’s Disease) typically have less of a Verbal Fluency Discrepancy, meaning less of an advantage of SVF over PVF. Those patients that have a PVF advantage over the SVF are highly suspected to convert to AD (Vaughan et al., 2016; Teng et al., 2013). Conversely, a primary impairment in the PVF task is often regarded as strong indicator for EF impairment and frontal-lobe degeneration indicating Dementia of the fronto-temporal type (Dubois et al., 2000). When it comes to AD there seems to be an overall agreement that the SVF is more sensitive for conversion from different stages (e.g. healthy, MCI AD) than PVF (Alegret et al., 2018; Murphy et al., 2006; Amieva et al., 2005).

Automatic qualitative analysis of both the SVF and the PVF using computational linguistics methods has shown very strong results in modelling strategy usage in this tasks (Lindsay et al., 2019; Tröger et al., 2019; Linz et al., 2017b) and ultimately also in automatically classifying between multiple cognitive disorders (König et al., 2018). However, research rarely takes into account both tasks or automatically calculates features to model the Verbal Fluency Discrepancy well-reported in clinical research.

Therefore the aim of this paper is to take advantage of the complementary diagnostic power of both VF tasks, combine this with previously established qualitative computational linguistics methods and prove such an approach’s overall quality for automatic classification in a traditionally very difficult applied machine learning scenario: Mild Cognitive Impairment vs. Healthy Controls.

## 2. Background

Traditionally, VF tasks are evaluated by counting all the relevant words produced in the given time frame, excluding repetitions. Although intrusions and repetitions have been investigated, state-of-the-art clinical evaluation of VF tasks centers around basic quantitative measures modelling neither qualitative aspects nor the temporal fine-grained resolution of a VF production.

### 2.1. Qualitative Evaluation of Verbal Fluency

Neuropsychological research investigated the quality of VF production early on by proposing a hierarchical set of rules to define qualitatively connected parts of a production (1997). The motivating rationale being that people do not produce words randomly but rather produce spurts of related words, or clusters. When a person has run out of easily accessible words, they intentionally navigate to a new associative field and exploit words from there, generally referred to as switching.

These early efforts propose for the SVF task a taxonomic approach with pre-defined semantic sub-categories (e.g. SVF on animals with subcategory farm animals or African animals). For the PVF task, a rule-based system is used to determine phonemic associations by manually defining criteria for phonemic similarity (Vonberg et al., 2014; Troyer et al., 1997); e.g. for PVF on the letter A, words are scored as associated/connected if they share common first letters like *arm* & *art*.

More recently multiple computational approaches have been proposed to model similar qualitative aspects within a VF performance. Approaches include inferring semantic associations in the SVF through distributional semantics (Linz et al., 2017a; Woods et al., 2016; Pakhomov and Hemmy, 2014), language models (Linz et al., 2018) or graph theory (Clark et al., 2016) and with data-driven approaches for multiple variations of edit distances in PVF (Lindsay et al., 2019).

Although automatic approaches to model qualitative aspects of VFs have been shown to be promising for both classification scenarios as well as classic inferential statistics experiments, they remain experimental and appear to be chosen subjectively if not arbitrarily.

Within this study, we will use automatic implementations of the rule-based early methods proposed by Troyer to keep results between the tasks comparable, as only this rule-based framework models qualitative aspects of SVF and PVF alike.

### 2.2. Temporal Evaluation of Verbal Fluency

While qualitative aspects of VF productions have been studied early on, temporal fine-grained modelling of such a production has been studied only recently. One main reason might be that temporal analysis of VF can only be performed if patients’ productions are recorded and transcribed. Tröger et al. (2019), suggests a temporal approach in which words that are said in close succession are considered to be in a cluster, regardless of semantic or phonemic motivation. The rationale behind it is that words which are—for whatever reason—associated in a person’s

semantic memory will be more accessible hence produced in faster succession.

While this has been used for the evaluation of the SVF task, there is—as of writing this paper—no research on the behavior of temporal clustering on the PVF task.

We choose to use this as one of our methods of feature generation as it does not require a semantically or phonemically motivated reason behind clustering and allows for an equal opportunity approach to both tasks.

### 2.3. Binning Approach to Verbal Fluency

Linz et al(2019) proposed a different temporal method for analyzing verbal fluency tasks where a one-minute speech sample is cut in to six 10-second bins. Features can then be calculated from each of the 10 second bins allowing for a finer resolution of the features over the task. In this paper, they looked at the word count, transition length and clustering features by bin. They found promising results using this technique in classifying between HC and MCI in Swedish subjects with the SVF Task, specifically for the word count of the last two intervals, 40-50 seconds and 50-60 seconds. These findings were supported by correlating different binning features with other trusted neuro-psychological tests such as the Boston Naming Test.

### 2.4. Automatic Classification of Verbal Fluency Tasks

Making use of novel qualitative as well as temporal features from VFs, recent work on automatic classification scenarios yields promising results. (Ryan, 2013) used logistic regression and a combination of SVF and PVF features to classify between healthy controls (HC) and MCI yielding an AUC of 0.76. (Linz et al., 2019) used temporal features extracted from SVF to classify between Swedish HC and MCI with the best result of an AUC of 0.72. Earlier in a classification experiment on French SVF data from HC and MCI (Linz et al., 2017a) achieved an F1 score of 0.77 by using qualitative semantic features.

	HC	MCI	<i>p</i>
N	8	19	
Sex (M/F)	8/0	12/7	-
Age (years)	71.50 (7.33)	75.32 (6.26)	0.18
Education (years)	9.75 (4.83)	10.53 (4.10)	0.67
MMSE (max 30)	29.25(0.89)	25.53 (3.31)	≤ 0.001

Table 1: Demographic information for the French population used in this analysis.

## 3. Methods

### 3.1. Data

The data used in this research was collected during the Dem@Care (Karakostas et al., 2017) and ELEMENT (Tröger et al., 2017) projects. Participants were recruited through the Memory Clinic located in Nice University Hospital at the Institute Claude Pompidou. Data was collected in the form of speech recordings via an automated recording application installed on a tablet computer. The recordings were manually transcribed in PRAAT (Boersma

Feature Name	Description
Word Count	The total number of animal words said in one minute, excluding repetitions
Mean Latency	Mean time (in seconds) elapsed since first utterance over all words
<i>Troyer Measures</i>	
Mean Troyer Cluster Size	Average number of animals in an SVF cluster over the entire sample
Number of Troyer Switches	the number of switches between Troyer clusters
<i>Temporal Measures</i>	
Mean Temporal Cluster length	Mean time (in seconds) spent inside a cluster.
Mean Temporal Cluster Coherence	Mean time (in seconds) spent between words inside clusters
Mean Temporal Cluster Size	the mean number of words inside a temporal cluster.
Number of Temporal Switches	the number of switches between temporal clusters
Mean Temporal Switch Coherence	Mean time (in seconds) between any two consecutive clusters.
<i>Bin Measures</i>	
Word Count by Bin	The number of words per 10 second bin
Transition Length by Bin	The average transition time in seconds between the end of one word and the onset of the next word by 10 second bin
<i>Difference Measures</i>	
Word Count	The total number of animal words said in one minute, excluding repetitions
Mean Latency	Mean time (in seconds) elapsed since first utterance over all words
Mean Troyer Cluster Size	Average number of animals in an SVF cluster over the entire sample
Number of Troyer Switches	the number of switches between Troyer clusters
Mean Temporal Cluster length	Mean time (in seconds) spent inside a cluster.
Mean Temporal Cluster Size	the mean number of words inside a temporal cluster.
Number of Temporal Switches	the number of switches between temporal clusters
Mean Temporal Switch Coherence	Mean time (in seconds) between any two consecutive clusters.
Word Count by Bin	Difference between SVF and PVF word count for respective 10s bin
Transition Length by Bin	SVF - PVF average transition in seconds between words for respective 10s bin

Table 2: The following features were extracted from the SVF and PVF task produced by the participants. For a more detailed explanation of how clusters are determined, please see Section 3.2.

and Weenink, 2009) according to the CHAT protocol (MacWhinney, 1991). Participants were asked to complete a battery of cognitive tests, including a 60 second semantic verbal fluency task—on the topic animals—and a 60 second phonemic verbal fluency task—for the letter category *F*. Demographics for the data used are displayed in 1 with significance testing between the populations. The MMSE (mini mental state examination) is a widely used test for cognitive performance where performance is measured on scale of 0 to 30 where anything below 25 is considered to be a sign of impairment. For this analysis four outliers were removed so that the maximum age considered was 85 years. One HC was removed for having an MMSE of 25, which would typically reflect some form of impairment.

### 3.2. Features

To investigate the diagnostic power of the SVF and PVF tasks, we designed three unique feature sets.

The first two are created by looking at each task individually; a *SVF feature set* and *PVF feature set*. An identical set of features were extracted from each task, according to how the task is evaluated. For example, the word count feature for the SVF is the number of animals said during the task, excluding repetitions and the for PVF it is the number of words starting with the letter *F* produced during the task, excluding repetitions. Word count, mean latency, Troyer measures, temporal measures and bin mea-

asures are extracted from each participant file for the single-task features sets and are described in the top-half of Table 2. A third feature set is created by combining the tasks by subtracting the PVF feature values from the corresponding SVF feature values of the same patient. This is referred to as the *difference features set*. For a detailed list of features and how they were produced see Table 2.

For the automatic computation of the Troyer clusters, SVF clusters are implemented according to the methodology in (Linz et al., 2017a) where a hierarchical set of predefined rules is used to determine semantically motivated clusters. Automatic Phonemic clusters, for the phonemic verbal fluency task are achieved by automating the phonetic rules proposed by Troyer as done in (Lindsay et al., 2019). Temporal clusters were computed according to Tröger et al. (2019). Binning measures for word count and transition length are calculated according to (Linz et al., 2019). As an additional temporal evaluation we computed mean latency, the average response latency for each word calculated by measuring the elapsed time since the onset of the first uttered word (Rohrer et al., 1995).

### 3.3. Oversampling with SMOTE

Due to a small and unbalanced data-set, the Synthetic Minority Oversampling Technique (SMOTE) is used to balance the HC and MCI population during the training phase of the classification experiments. This technique oversam-

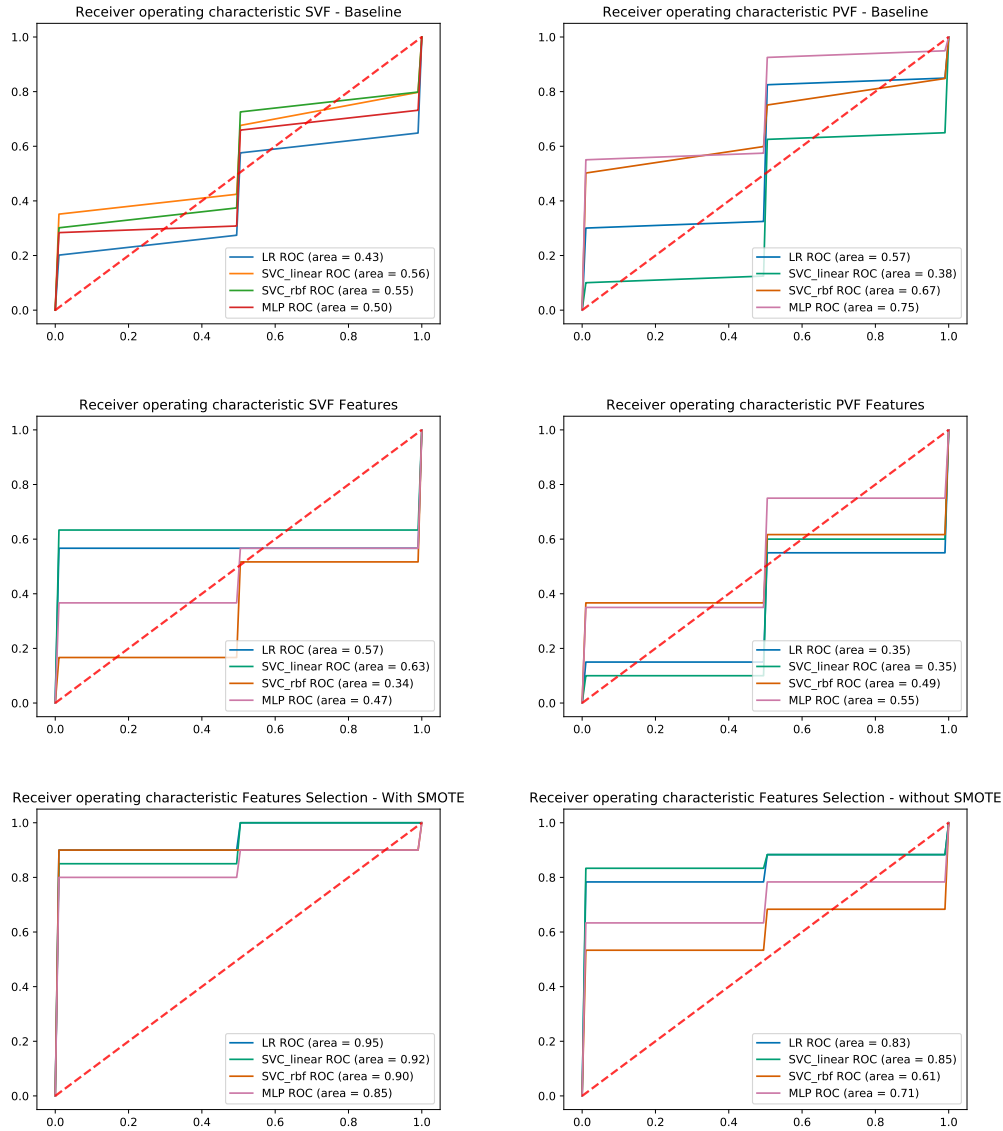


Figure 1: ROC Curves and corresponding AUC values for the HC vs. MCI classification experiments. Models are indicated by color. The red dashed line represents chance performance.

ples the minority class—in this case, healthy controls—by creating synthetic healthy samples so that the data becomes balanced. Synthetic samples are created by randomly choosing an existing example and drawing a line to other similarly existing samples within the features space. New synthetic samples are then randomly chosen points from this line (Bowyer et al., 2011). SMOTE is applied during the cross-fold validation, after splitting the fold data into training and testing sets, SMOTE is applied only on the training data. The testing split is not augmented. Therefore, testing is only done on real samples.

### 3.4. Classification Experiments

To consider the validity of this approach, a series of machine learning experiments are conducted with the different proposed features sets, simulating a screening scenario: automatically differentiating between healthy controls and patients with mild cognitive impairment. The focus of this paper is on the application of looking at the feasibility of

combining PVF and SVF for early diagnosis of Dementia.

Six experimental scenarios were created from the features described in Section 3.2.: 1) an SVF Clinical Baseline where only the SVF word count is considered, 2) a PVF Clinical Baseline where only the PVF word count is considered, 3) the SVF feature set, 4) the PVF feature set, 5) the SVF, PVF and DIFF feature set and 6) the SVF, PVF and DIFF feature set without SMOTE. To obtain a more comprehensive picture, four machine learning approaches are considered:

- A Logistic Regression (LR) model is created where error is minimized by least square errors, what is commonly referred to as the L2 loss.
- A Multilayer Perceptron (MLP) model is used with the Limited-memory BFGS (lbfgs) solver and logistic activation function. The alpha parameter is set at 0.1 and an adaptive learning rate is used.

- A Support Vector Machine Classification model is also created with a linear (SVC-linear) kernel.
- A Support Vector machine Classification with a radial basis function (SVC-rbf) kernel.

For each experiment, 5-fold cross validation is used All models are created simultaneously so that the same split of the data in each fold is used to train and test each of the four models. No parameter optimization is used. For scenarios 5) the SVF, PVF and Difference feature set with SMOTE and 6) the SVF, PVF and DIFF feature set without SMOTE, features selection is used in the training phases where and independent t-test is used to determine the significance of the features between the groups. Features with a p-value greater than 0.05 are discarded and the remaining features are used to train. The classification models are created using the scikit-learn library in Python3 (Pedregosa et al., 2011).

## 4. Results

### 4.1. Classification Results

Results from the classification experiment are displayed in Table 3. For evaluation, accuracy, sensitivity, precision, F1 score and Area Under the Receiver Operator Curve (AUC) are provided. The mean score from the 5-fold cross validation is given as well as the standard deviation in parentheses. Each feature set is displayed with the results from each model described in Section 3.4.. Results are visualized with receiver operator characteristic (ROC) curves in Figure 1, where a larger area under the curve (AUC) indicates that the model is better at differentiating between HC and MCI.

During the feature selection process, the features that were chosen based on their significance were *PVF word count<sub>40-50</sub>*, *SVF word count<sub>40-50</sub>*, *SVF transition length<sub>40-50</sub>*, *DIFF word count<sub>10-20</sub>*.

From the clinical baseline models, where just word count is used, SVF showed consistent AUCs of roughly 0.5 across models, with the highest AUC of 0.56 coming from SVC-Linear. For PVF, AUC scores varied in the baseline from a low of 0.38 by the SVC-Linear and the highest AUC of 0.75 coming from the MLP. Using all the automated SVF features, SVF improved from the baseline, achieving its highest score with SVC-Linear at 0.63. An increase of 0.07 from the SVF baseline. PVF decreased with the additional automated features from the word count baseline with its highest AUC reaching 0.55. The best results are found when combining the SVF, PVF and DIFF features. The highest AUC of 0.95 is achieved by LR. The lowest AUC of 0.85 is achieved by the MLP. The Accuracy of these models is maintained between 70 and 80%.

We consider this scenario without SMOTE to see how oversampling might affect the training. There is a slight dip in performance. However it exceeds all other SVF and PVF models. The highest AUC of 0.86 is achieved by SVC-Linear at 0.86. The lowest AUC is found at 0.61 by the SVC-RBF. However, the accuracy is, on average, higher

than with SMOTE. The average accuracy across models with SMOTE is 74.8% but without SMOTE it is 75.5%.

## 5. Discussion

Looking at the results from the classification models, there is improvement from the clinical baseline of word count from both the SVF and PVF task in comparison to the all features model both with and without SMOTE, which hinges on the additional computational measures.

The classification for both the SVF-baseline (AUC=0.56) and all SVF features (AUC=0.63) seems low. Previous papers reported AUCS of over 0.7 using similar features and models to distinguish between HC and MCI (Linz et al., 2019)(Linz et al., 2017b). One difference that may have lead to this result is using a hierarchical predefined list to determine clusters instead of using an automated approach, such as clustering using semantic word embeddings. We also chalk this low result up to the relative size of the data set.

However, for SVF there is at least some improvement over the baseline using the additional computational measures. This is not mirrored by the PVF task where the baseline (AUC = 0.75) is better than the additional features (AUC = 0.55). PVF lacks the foundation of research in computational measures. From a feature standpoint, previous methods that have shown to be beneficial for evaluating the SVF task seem to transfer to the PVF task (e.g. clustering, binning, ect). Future work should look at the underlying production strategies and cognitive processes engaged, during the PVF, similar to the work that has been done on SVF, in order to improve its classification as a standalone task.

An interesting finding from the feature selection is that features from each data set were used to achieve the best classification result; PVF word *count<sub>40-50</sub>*, SVF word *count<sub>40-50</sub>*, SVF transition *length<sub>40-50</sub>*, DIFF word *count<sub>10-20</sub>*. This highlights the need for diverse features even in a small data setting. (Linz et al., 2019) also found similar results at the 40-50s bin for SVF in an early diagnosis scenario for dementia. They found that SVF word count in the 40-50s bin correlated positively with scores for other neuro-psychological tests that measure vocabulary, namely the Boston Naming Test and WAIS similarity task. It is interesting to see that this finding is repeated in the PVF task. This is something that should be investigated for the PVF task with a larger data set. It should also be considered for further investigations of underlying cognitive processes. This result also highlights the benefits of using binning as a qualitative temporal analysis method of the verbal fluency tasks.

While some of the machine learning classification results presented are quite accurate, this may be due to the synthetic data augmentation technique, SMOTE. While clinical data tends to be relatively noisy, the synthetic data is probably a cleaner data set than real life conditions. To test this, we need more data and a balanced data set to confirm that the classification results presented here hold in real-world testing conditions. While we do expect slightly worse results without SMOTE—as shown in

Model	Accuracy	Sensitivity	Specificity	Precision	F1	AUC
Semantic Verbal Fluency - Clinical Baseline						
LR	0.48(0.11)	0.53(0.17)	0.30(0.24)	0.65(0.08)	0.58(0.14)	0.42(0.26)
SVC - Linear	0.53(0.17)	0.62(0.17)	0.30(0.24)	0.68(0.10)	0.64(0.14)	<b>0.56(0.27)</b>
SVC - RBF	0.43(0.14)	0.47(0.17)	0.30(0.24)	0.62(0.10)	0.53(0.15)	0.55(0.23)
MLP	0.56(0.13)	0.65(0.25)	0.40(0.37)	0.75(0.14)	0.65(0.16)	0.50(0.10)
Phonemic Verbal Fluency - Clinical Baseline						
LR	0.49(0.24)	0.45(0.40)	0.70(0.24)	0.44(0.36)	0.44(0.37)	0.57(0.29)
SVC - Linear	0.68(0.19)	0.77(0.29)	0.50(0.32)	0.81(0.11)	0.75(0.17)	0.38(0.27)
SVC - RBF	0.86(0.07)	0.95(0.10)	0.60(0.37)	0.88(0.10)	0.90(0.05)	0.68(0.26)
MLP	0.79(0.14)	0.85(0.12)	0.60(0.37)	0.86(0.12)	0.85(0.09)	<b>0.75(0.21)</b>
Semantic Verbal Fluency Features						
LR	0.55(0.15)	0.58(0.25)	0.40(0.37)	0.76(0.13)	0.62(0.16)	0.57(0.28)
SVC - Linear	0.55(0.10)	0.58(0.25)	0.40(0.49)	0.80(0.16)	0.62(0.13)	<b>0.63(0.25)</b>
SVC - RBF	0.51(0.17)	0.58(0.19)	0.30(0.24)	0.67(0.09)	0.61(0.16)	0.34(0.24)
MLP	0.48(0.11)	0.63(0.19)	0.10(0.20)	0.63(0.07)	0.62(0.10)	0.47(0.32)
Phonemic Verbal Fluency Features						
LR	0.56(0.13)	0.65(0.30)	0.40(0.37)	0.75(0.13)	0.64(0.17)	0.35(0.22)
SVC - Linear	0.55(0.10)	0.63(0.25)	0.40(0.37)	0.75(0.13)	0.64(0.14)	0.35(0.18)
SVC - RBF	0.66(0.18)	0.70(0.29)	0.60(0.37)	0.84(0.13)	0.71(0.20)	0.49(0.31)
MLP	0.56(0.13)	0.65(0.30)	0.40(0.37)	0.75(0.13)	0.64(0.17)	<b>0.55(0.33)</b>
All Features with Feature Selection - With SMOTE						
LR	0.77(0.14)	0.78(0.19)	0.80(0.24)	0.91(0.11)	0.82(0.11)	<b>0.95(0.10)</b>
SVC - Linear	0.71(0.12)	0.73(0.16)	0.70(0.24)	0.84(0.13)	0.77(0.11)	0.93(0.10)
SVC - RBF	0.79(0.18)	0.85(0.20)	0.70(0.40)	0.88(0.15)	0.84(0.13)	0.90(0.12)
MLP	0.72(0.18)	0.80(0.19)	0.50(0.32)	0.79(0.11)	0.79(0.14)	0.85(0.12)
All Features with Feature Selection - without SMOTE						
LR	0.78(0.06)	0.80(0.10)	0.70(0.40)	0.90(0.12)	0.84(0.04)	0.83(0.14)
SVC - Linear	0.76(0.13)	0.83(0.14)	0.60(0.37)	0.85(0.13)	0.83(0.08)	<b>0.86(0.13)</b>
SVC - RBF	0.74(0.07)	0.95(0.10)	0.20(0.24)	0.75(0.05)	0.84(0.05)	0.61(0.24)
MLP	0.74(0.09)	0.90(0.12)	0.30(0.40)	0.79(0.11)	0.83(0.05)	0.71(0.12)

Table 3: This table contains results from the classification experiments. All Features is the combination of SVF, PVF and difference features. For evaluation, accuracy, sensitivity, precision, F1 score and Area Under the Receiver Operator Curve (AUC) are provided. The mean of the 10-fold validation is given. The standard deviation is given in parentheses. Highest Accuracy and AUC scores are emphasized in bold font. All models use SMOTE except for the final experiment which is labeled *All Features - Without SMOTE*.

the All Features without SMOTE classification experiment (SVC-Linear AUC=0.86)—we still expect improved results with more authentic data. It is worth noting that SMOTE is not being used to increase the accuracy or AUC of the model directly, but is being used to improve the training data, which could indirectly influence the models’ behavior.

## 6. Conclusion

In this paper, we explored using previously automated qualitative analysis techniques from the SVF—semantic and temporal clustering as well as temporal binning—on the PVF tasks with promising results. Moreover, as clinical research suggests, we present an approach to fuse both tasks in calculating specific difference features harnessing the so-called Verbal Fluency Discrepancy in AD and its precursor stage MCI. The features generated from both of these tasks as well as the development of multi-task joint difference features lead to improved classification for early detection of Dementia symptoms and are verified by multiple classi-

fiers, with and without synthetic data augmentation to balance a small clinical data set. While the results are promising, this paper setup a pipeline for the feasibility of creating classification experiments with multiple verbal fluency tasks. This work should be reiterated in other languages as well as on larger data sets to confirm the suggested conclusions.

## 7. Bibliographical References

- Alegret, M., Pereto, M., Perez, A., Valero, S., Espinosa, A., Ortega, G., Hernandez, I., Mauleón, A., Rosende-Roca, M., Vargas, L., et al. (2018). The role of verb fluency in the detection of early cognitive impairment in alzheimer’s disease. *Journal of Alzheimer’s Disease*, 62(2):611–619.
- Amieva, H., Jacqmin-Gadda, H., Orgogozo, J.-M., Le Carret, N., Helmer, C., Letenneur, L., Barberger-Gateau, P., Fabrigoule, C., and Dartigues, J.-F. (2005). The 9 year cognitive decline before dementia of the alzheimer

- type: a prospective population-based study. *Brain*, 128(5):1093–1101.
- Boersma, P. and Weenink, D. (2009). Praat: doing phonetics by computer (version 5.1.13).
- Bowyer, K. W., Chawla, N. V., Hall, L. O., and Kegelmeyer, W. P. (2011). SMOTE: synthetic minority over-sampling technique. *CoRR*, abs/1106.1813.
- Clark, D. G., McLaughlin, P. M., Woo, E., Hwang, K., Hurtz, S., Ramirez, L., Eastman, J., Dukes, R. M., Kapur, P., DeRamus, T. P., and Apostolova, L. G. (2016). Novel verbal fluency scores and structural brain imaging for prediction of cognitive outcome in mild cognitive impairment. *Alzheimers Dement (Amst)*, 2:113–122.
- Dubois, B., Slachevsky, A., Litvan, I., and Pillon, B. (2000). The fab. *Neurology*, 55(11):1621–1626.
- Karakostas, A., Briassouli, A., Avgerinakis, K., Kompatsiaris, I., and Tsolaki, M. (2017). The dem@care experiments and datasets: a technical report. *CoRR*, abs/1701.01142.
- König, A., Linz, N., Tröger, J., Wolters, M., Alexandersson, J., and Robert, P. (2018). Fully automatic analysis of semantic verbal fluency performance for the assessment of cognitive decline. *Dementia and Geriatric Cognitive Disorders*. Accepted.
- Lindsay, H., Linz, N., Tröger, J., and Alexandersson, J. (2019). Automatic data-driven approaches for evaluating the phonemic verbal fluency task with healthy adults. In *ICNLSP*.
- Linz, N., Tröger, J., Alexandersson, J., and König, A. (2017a). Using Neural Word Embeddings in the Analysis of the Clinical Semantic Verbal Fluency Task. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS)*.
- Linz, N., Tröger, J., Alexandersson, J., Wolters, M., König, A., and Robert, P. (2017b). Predicting dementia screening and staging scores from semantic verbal fluency performance. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 719–728, Nov.
- Linz, N., Tröger, J., Lindsay, H., König, A., Robert, P., Peter, J., and Alexandersson, J. (2018). Language modelling for the clinical semantic verbal fluency task. In Dimitrios Kokkinakis, editor, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may. European Language Resources Association (ELRA).
- Linz, N., Lundholm Fors, K., Lindsay, H., Eckerström, M., Alexandersson, J., and Kokkinakis, D. (2019). Temporal analysis of the semantic verbal fluency task in persons with subjective and mild cognitive impairment. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 103–113, Minneapolis, Minnesota, June. Association for Computational Linguistics.
- MacWhinney, B. (1991). *The CHILDES project: Tools for analyzing talk*. Lawrence Erlbaum Associates, Inc.
- Murphy, K. J., Rich, J. B., and Troyer, A. K. (2006). Verbal fluency patterns in amnesic mild cognitive impairment are characteristic of alzheimer’s type dementia. *Journal of the International Neuropsychological Society*, 12(4):570–574.
- Pakhomov, S. V. and Hemmy, L. S. (2014). A computational linguistic measure of clustering behavior on semantic verbal fluency task predicts risk of future dementia in the nun study. *Cortex*, 55:97–106.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Rohrer, D., Wixted, J. T., Salmon, D. P., and Butters, N. (1995). Retrieval from Semantic Memory and its Implications for Alzheimer’s Disease. *J Exp Psychol Learn Mem Cogn*, 21(5):1127–1139.
- Ryan, J. (2013). *A System for Computerized Analysis of Verbal Fluency Tests*. Ph.D. thesis, 06.
- Teng, E., Leone-Friedman, J., Lee, G. J., Woo, S., Apostolova, L. G., Harrell, S., Ringman, J. M., and Lu, P. H. (2013). Similar Verbal Fluency Patterns in Amnesic Mild Cognitive Impairment and Alzheimer’s Disease. *Archives of Clinical Neuropsychology*, 28(5):400–410, 06.
- Tröger, J., Linz, N., Alexandersson, J., König, A., and Robert, P. (2017). Automated Speech-based Screening for Alzheimer’s Disease in a Care Service Scenario. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*.
- Troyer, A. K., Moscovitch, M., and Winocur, G. (1997). Clustering and Switching as Two Components of Verbal Fluency: Evidence From Younger and Older Healthy Adults. *Neuropsychology*, 11(1):138–146.
- Tröger, J., Linz, N., König, A., Robert, P., Alexandersson, J., Peter, J., and Kray, J. (2019). Exploitation vs. exploration—computational temporal and semantic analysis explains semantic verbal fluency impairment in alzheimer’s disease. *Neuropsychologia*, 131, 05.
- Vaughan, R. M., Coen, R. F., Kenny, R., and Lawlor, B. A. (2016). Preservation of the semantic verbal fluency advantage in a large population-based sample: Normative data from the tilda study. *Journal of the International Neuropsychological Society*, 22(5):570–576.
- Vonberg, I., Ehlen, F., Fromm, O., and Klostermann, F. (2014). The absoluteness of semantic processing: Lessons from the analysis of temporal clusters in phonemic verbal fluency. 9:e115846, 12.
- Woods, D. L., Wyma, J. M., Herron, T. J., and Yund, E. W. (2016). Computerized Analysis of Verbal Fluency: Normative Data and the Effects of Repeated Testing, Simulated Malingering, and Traumatic Brain Injury. *PLOS ONE*, 11(12):1–37.



# Toward Characterizing the Language of Adults with Autism in Collaborative Discourse

Christine Yang<sup>†</sup>, Emily Prud'hommeaux<sup>†</sup>, Laura B. Silverman<sup>‡</sup>, Allison Canfield\*

<sup>†</sup>Boston College, Chestnut Hill MA, USA,

<sup>‡</sup>University of Rochester, Rochester NY, USA

\*Upstate Medical Center, Syracuse NY, USA

{yangael, prudhome}@bc.edu, laura\_silverman@urmc.rochester.edu

## Abstract

Autism spectrum disorder (ASD) is a neurodevelopmental condition associated with life-long deficits in communication that can impact both personal and professional well-being. Although the linguistic features associated with these deficits are routinely observed in clinical settings, they are difficult to quantify. In this paper, we present a growing dataset of conversations between high-functioning adults with ASD and their neurotypical conversational partners as they complete a collaborative task. We compare the linguistic characteristics of the two groups using both manual annotations and computational linguistic features extracted from these conversations. Our results indicate that there are quantifiable differences in the language use of adults with ASD in collaborative discourse scenarios, demonstrating the promise of our methods and dataset.

**Keywords:** autism spectrum disorder, discourse, pragmatics, spoken language analysis, corpus analysis, speech acts

## 1. Introduction

Autism spectrum disorder (ASD) is a neurodevelopmental condition associated with life-long deficits in communication and social engagement. Among these deficits is impaired pragmatic expression, or the inappropriate use of language in a given context (Kanner, 1943; Lord and Paul, 1997; Young et al., 2005; Simmons et al., 2014). Because of the pragmatic difficulties they experience, individuals with ASD face challenges in establishing interpersonal relationships, maintaining satisfactory employment, and achieving independence (Mok et al., 2014; Whitehouse et al., 2009; Hendricks, 2010). Researchers do not agree, however, on precisely what functions are impaired, particularly in high-functioning adults. Analyzing spontaneous spoken language is an effective way to reveal these impairments, but there has been relatively little work on either manually annotating or computationally analyzing spontaneous language data from adults with ASD. As a result, there are no publicly available conversational spoken language datasets produced by adults with ASD.

In this paper, we describe a growing dataset of transcribed conversations between high-functioning adults with ASD or typical development (TD) and their neurotypical conversational partners as they work together to navigate from one location to another on a shared map. Although the number of study participants whose collaborative conversations has been transcribed thus far is modest, the data collection project that is the source of these conversations is ongoing and will include, within the next 18 months, 60 to 75 participants. The dataset includes conversations from two additional collaborative tasks and as well as spoken responses to a variety of widely used clinical instruments. We will be making the transcripts, as well as the manual annotations we have created, available to researchers who can demonstrate that they have completed their institution's human subjects

training in the hopes that the data will reveal new and useful information about the strengths and weaknesses in pragmatic expression associated with ASD.

Here we present the results of both manual and automated computational analyses of the data collected so far. Our findings suggests that there are observable and quantifiable differences between adults with ASD and those with typical development on several discourse-level pragmatic dimensions. These results underscore the importance of examining spontaneous conversational speech in adults with ASD and point to the promise of automated computational approaches for clinical language analysis.

## 2. Related Work

Atypical language has been observed in verbal individuals with autism since the disorder was first described (Kanner, 1943) and continues to serve as a diagnostic criterion in many widely used instruments for diagnosing autism (Lord et al., 2002; Rutter et al., 2003). Atypical use of language for a given context, known as pragmatic expression, seems to be universally affected in autism, even in the absence of structural language impairments in syntax, morphology, and phonology (Eales, 1993; Landa, 2000; Young et al., 2005; Simmons et al., 2014). In high-functioning individuals with autism, impaired pragmatic expression is associated with challenging behaviors (Ketelaars et al., 2010), difficulty developing relationships (Whitehouse et al., 2009), and struggles in maintaining employment (Hendricks, 2010).

The most promising methods for pinpointing the pragmatic features that characterize autism rely on careful manual annotation of transcripts of spontaneous spoken language (Volden and Lord, 1991; Bishop et al., 2000; Adams, 2002; Gorman et al., 2016; Canfield et al., 2016). Carrying out

Dx	PIQ	Age
TD ( $n=5$ )	103 (10.5)	22y5m (4y3m)
ASD ( $n=9$ )	106 (8.4)	19y11m (1y9m)

Table 1: Summary demographic statistics of our current set of participants: mean (s.d.).

complex annotations schema, however, requires training and expertise, making these methods are impractical to deploy. There has been relatively little work in applying computational methods for identifying these sorts of linguistics features in the language of individuals with ASD, and this work has focused exclusively on the language of children and language produced in a semi-structured context (Prud’hommeaux et al., 2014; Losh and Gordon, 2014; Parish-Morris et al., 2016; Goodkind et al., 2018).

The language resource and accompanying analysis presented here makes several novel contributions. First, this language data is produced by adults, a subgroup of the ASD population that is both understudied and underserved. Second, the dataset consists entirely of spontaneous conversations in a restricted semantic domain. Third, the dataset has been manually annotated to indicate the category of speech act for each turn and a numeric rating on several scales, including politeness, uncertainty, and informativeness.

### 3. Data Collection

#### 3.1. Spoken Language Data

As part of a project investigating differences in pragmatic expression in adults with ASD, we are collecting spoken language data from high-functioning adult participants with ASD and with typical development (TD). Participants with ASD must meet criteria for a diagnosis of ASD on the Autism Diagnostic Observation Schedule (ADOS) (Lord et al., 2002), as well as the following basic eligibility criteria: (1) full-scale IQ (PIQ) > 80; (2) verbal IQ (VIQ) > 80; (3) monolingual American English speaking; and (4) no history of language impairment, auditory processing disorder, or hearing difficulty. Neurotypical participants are selected in order to match the experimental participants on age, VIQ, PIQ, gender, and ethnicity. Because our data collection is in progress, the participants analyzed here may not yet be matched on all dimensions. Table 1 presents mean values for age, full-scale IQ for our 9 participants with ASD and 5 participants with TD.

Each participant, whether with ASD or TD, is paired with a neurotypical conversational partner, with whom they perform several collaborative tasks that require verbal communication and deliberation. In one task, each person is given a map of the same place, but with slight differences in the place names and the location of obstacles referred to as “road blocks” (Anderson et al., 1991). Each map is also labeled with a mark to show where the conversational partner is located. The participant’s task is to give

directions to the conversational partner to get them to their position on the map. Currently, we have collected recorded conversations from 14 pairs of participants, with 5 participants in the TD group and 9 participants in the ASD group.

After the spoken data is collected, the recordings are transcribed using Praat (Boersma, 2001). All filler words, discourse markers, and words or sounds of affirmation, negation, or exclamation are included in the transcripts, as these serve as important tools for expression and may be informative for pragmatic analysis. Transcriptions also included annotations for sounds effects or onomatopoeia, partial or interrupted words, and unfilled pauses within an utterance. Utterances are segmented using the concept of the *C-unit*, which is formally defined as “an independent clause with its modifiers” (Loban, 1976). It includes the main clause and all subordinate clauses, and cannot be further segmented without losing its essential meaning. It does not have to be a complete syntactical sentence, and may comprise of a single coordinate clause (using coordinating conjunctions “and”, “so”, “then”, etc.), but not a single subordinate clause (using subordinating conjunctions “because”, “if”, “when”, etc.). Each utterance is punctuated in the transcript with one of several punctuation marks for exclamations, questions, regular statements, abandoned utterances, and interrupted utterances. Our corpus of 14 transcripts currently consists of about 4,463 utterances in total, with 3,019 utterances in the TD group and 1,444 in the ASD group. Of these, we have 2,240 utterances from the participants which we use for the analyses that follow.

#### 3.2. Linguistic Features

For each utterance, we gathered several pragmatic features which we believed could be significant in illuminating the differences between adults with ASD and adults with TD. In our selection of potentially significant linguistic features, we explored both manually annotated features as well as computationally derived features generated by existing models and toolkits, described in the subsections below. The manual annotations investigated potential pragmatic differences identifiable by human observers using a set annotation guidelines, under the assumption that a pragmatic feature perceptible to an annotator would also be perceptible to a conversational partner and thus have an effect in real-world communication and pragmatics. We also use existing computational tools to predict further pragmatic features for each utterance. The ability of the automated features to capture meaningful conclusions is, of course, dependent on the model and corpus used to generate the feature ratings, but it is still worth investigating these features to see whether they may point to some linguistic differences between the two groups that are not captured by the manual annotations. The predicted features are also much easier and less time-consuming to acquire and may thus help us determine which additional features might be worthwhile for future exploration and annotation.

### 3.2.1. Manually Annotated Features

Two human annotators were assigned to annotate each utterance with a numerical score for politeness, uncertainty, and information content. Each category was originally rated on a discrete scale of 1 to 5, but it was later collapsed to a scale of 1 to 3 as the smaller scale helped improve inter-annotator agreement. Each utterance was treated independently and rated without consideration for the context surrounding it, as the eventual goal is to potentially train a model that can assign these ratings automatically on an utterance by utterance level. Therefore, identical utterances in different contexts were given the same feature ratings. The annotated feature categories are defined as follows:

**Politeness:** The *politeness* rating is a measure of how positive, agreeable, and non-demanding an utterance is. A most polite utterance shows high positivity and willingness to compromise or admit wrongdoing. An utterance with a politeness score of 2 was given to neutral statements, which included direct questions (“where are you?”), objective observations (“the house is red”), and commands phrased indirectly (“then you wanna go left”, “then you’re gonna go left”). A high politeness rating of 3 was given to utterances which included positive or affirmative words (“great”, “I agree”, “true”), acknowledged the speaker’s own mistakes, or contained distinct politeness markers (“please”, “thank you”). Commands phrased as conditionals (“if you wanna make a left”), suggestions (“how about you go left”), or directed to both of them using the first person plural (“then we need to go left”) had a score of 3. Requests using a modal (“could you tell me”) were also given a rating of 3. A low politeness score of 1 was given to utterances which contained negative comments or expressed frustration (“how the heck am I supposed to say this?”), criticized the other person, or directly accused the other person of being wrong. Commands phrased as imperatives (“go left”) or as necessity for the other person (“you have to go left”, “I need you to go left”) were also given a score of 1.

**Uncertainty:** The *uncertainty* rating is a measure of how uncertain the speaker is about a fact or about the accuracy of their utterance. An utterance with an uncertainty rating of 1 showed no clear signs of uncertainty, while a rating of 2 indicated some hesitation (filler words, pauses), hedging (“maybe”, “might be”), or qualification to the statement (“if I’m reading this correctly”). Polar questions (“Is it red?”) and words or phrases intended as questions for confirmation (“The one by the tree?”) were also assigned an uncertainty rating of 2. Questions expecting a one word or phrase answer (“What color is the building?”) and questions expecting longer explanations (“How do I get there from here?”) were given the highest rating of 3. Directly stating “I don’t know” or “I’m confused” was also given a high uncertainty rating of 3.

**Information Content:** The *information content* rating is a measure of the quantity and specificity of the information words contained in the utterance. Utterances containing no information at all and utterances containing

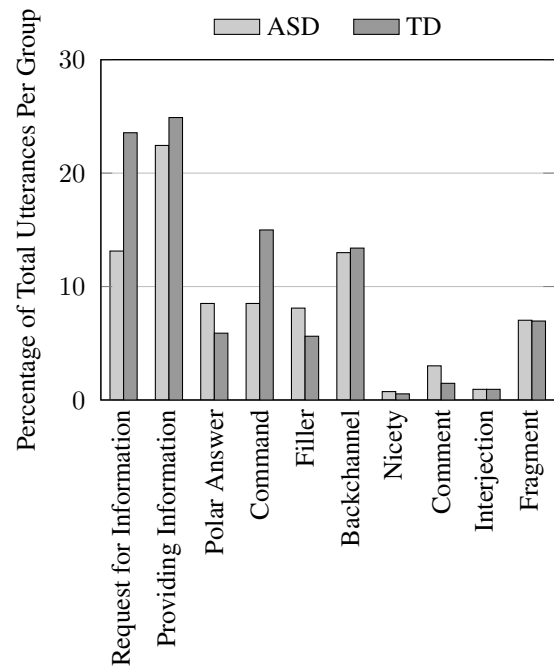


Figure 1: Speech act distribution per group.

some vague pronouns or polar answers (“yes”, “I don’t have it”, “that one”) were given a rating of 1. A score of 3 was given to utterances containing directional words (“left”, “north”, “down”) or general object words that could refer to multiple items on the map (“bird”, “red roof”, “road block”). Utterances with the highest rating of 3 contained proper nouns and specific place names (“Hawk\_Meadow”, “compost site”) or elaborate descriptions of objects or places that could only refer to a specific location on the map (“a big red house with two windows on the left side and four windows in the middle”).

**Speech Acts:** Along with these features, the annotators were also asked to assign a speech act to each utterance. The set of speech acts used to annotate this specific dataset is defined in Table 2.

Between the two annotators, all 14 transcripts were annotated for the features and speech acts detailed above, with 8 transcripts having annotations from both annotators. To determine the inter-annotator agreement, we calculated the percentage of agreement and Cohen’s *kappa* (Cohen, 1968) for the utterances that had been annotated by both annotators. Results are shown in Table 3.

The agreement for each category was above 80%, and the kappa scores for uncertainty, information content, and speech act were all in the substantial range of 0.61 to 0.80, as defined by Cohen (Cohen, 1968). The politeness feature had a high agreement but a lower kappa score of 0.54, likely because most of the utterances had a neutral politeness rating, which was true under both the original 5-point scale and the collapsed 3-point scale. This was likely due to the fact that the majority of utterances did not attempt to use any identifiable politeness strategies.

Speech Act	Description
Request for Information	Any request for information, clarification, or confirmation. May take the form of a question, or of a statement or phrase intended to be a question (“You want me to keep going?”, “the one by the tree?”)
Providing Information	Answering a request for information, or providing information unprompted.
Polar Answer	Answering “yes” or “no” to a polar question.
Command	An utterance that gives instruction or direction to the other person whether in the form of an imperative (“go left.”), a suggestion (“how about you go left”), a hypothetical (“if you wanna make a left”), or a question (“do you wanna go left?”). It may be instructing the person on where to go on the map, or on how to act or strategize (“hold on”, “how about you tell me where you are first”).
Filler	Filler words or phrases used to fill pauses in the conversation (“hm”, “anyways”, “okay so”).
Backchannel	An utterance that indicates the speaker is listening and understanding what the other person is saying (“okay”, “mm-hm”, “gotcha”, “sounds good”).
Nicety	Utterances that serve to maintain a polite and collaborative conversation, such as apologizing, expressing gratitude, or reassuring the conversational partner.
Comment	An utterance that contains commentary or an opinion on the task, such as explaining the speaker’s own actions or discussing the best strategy to take.
Interjection	Short exclamations or interjections such as “ah”, “oops”, “yay”, “wow”, “ew”, “awesome”, etc.
Fragment	Short abandoned or interrupted utterances that are too incomplete to classify as any other speech act.

Table 2: The set of speech acts used in annotation and their descriptions with some examples.

Feature	Agreement	Kappa ( $\kappa$ ) Score
Politeness	89.83%	0.544
Uncertainty	84.18%	0.691
Information Content	85.82%	0.758
Speech Act	81.19%	0.777

Table 3: Inter-annotator Agreement

### 3.2.2. Automated Computational Features

In addition to these manually annotated features, we also explored several computationally derived features extracted using existing toolkits and models used previously to characterize conversations in collaborative software development (Meyers et al., 2018). In particular, we extracted scores for politeness, uncertainty, formality, informativeness, and implicature. The *politeness* classifier uses an SVM model trained on over 10,000 annotated requests from online forums. It uses the Stanford CoreNLP software to generate dependency parses for preprocessing and assigns each utterance a politeness rating on a continuous scale from 0 to 1, with 1 being the most polite. The *uncertainty* classifier uses a logistic regression model trained on

the Szeged Uncertainty Corpus (Vincze, 2014) and assigns each utterance a binary classification of either certain or uncertain. In this package, an uncertain utterance is defined as one for which the “truth value or reliability cannot be determined due to lack of information” (Vincze, 2014). The *squinky* package (Meyers et al., 2018) uses a logistic regression model trained on a corpus of over 7,000 annotated sentences and rates each utterance on formality, informativeness, and implicature on a scale of 0 to 1, with 1 being the most formal, informative, and implicative respectively. Formality is a linguistic strategy employed to effectively convey as much information as possible while adhering to Grice’s maxims. The informativeness scale is related to the concept of *term informativeness*, and corresponds to how clearly and how directly the intended meaning is communicated. Implicature is a measure of the amount of missing or implied information in an utterance. We refer the reader to Meyers et al. (2018) and Vincze (2014) for further details.

## 4. Data Analysis

The frequency of speech acts in each group, shown in Figure 1, generally shows a similar distribution across diagnostic groups, with an increased usage in the TD group of requests for information, commands, and providing information. The percentage of speech acts in the TD group that are requests for information is over 10% higher than that of the ASD group. The fact that the TD group

Feature	ASD Average Rating	TD Average Rating
<b>Manually Annotated Features (scale of 1 to 3)</b>		
Politeness***	<b>2.02</b>	1.93
Uncertainty*	1.43	<b>1.48</b>
Information Content***	1.54	<b>1.77</b>
<b>Automated Computational Features (scale of 0 to 1)</b>		
Politeness	<b>0.448</b>	0.445
Uncertainty	0.954	<b>0.958</b>
Formality	<b>0.016</b>	0.014
Informativeness	<b>0.105</b>	0.089
Implicature***	0.35	<b>0.414</b>

Table 4: Average manual ratings and automated scores for pragmatic features. One asterisk indicates a significant difference between the two groups ( $p < 0.05$ ). Three asterisks (\*\*\*) indicates a highly significant difference between the two groups ( $p < 0.0001$ ).

made greater use of speech acts which often include more information content might also contribute to the fact that the TD utterances were rated more highly for information content on average.

The two groups of participants generally had similar mean values for the annotated features, as shown in Table 4. For the manually annotated features, the ASD group was slightly more polite while the TD group contained more information content overall. The automated computational features also appeared to have very close averages between the two groups, with the exception of the implicature feature, for which the ASD participants’ utterances scored noticeably higher.

To determine if the differences between the two groups were significant, we performed a two-tailed t-test over all the utterances of each participant group, with the results displayed in Table 4. This significance testing revealed that all three manually annotated features were significantly different across the two groups, with the ASD group having more polite utterances and the TD group showing more uncertainty and information content. The t-test also showed that there were generally more significant differences in the manually annotated features than in the automated features, with the exception of the automated implicature score which had a highly significant difference between the two diagnostic groups.

We also note that the two groups might employ different pragmatic strategies for different speech acts. To examine the linguistic features of the utterances for each individual speech act, we calculated average linguistic feature score

for each speech act. The results for some speech acts of interest are displayed in Table 5. We again performed a t-test on this data, and from the results we can see that while there are no significant differences between the two groups when requesting information, there are significant differences in levels of politeness and information content when providing information, and in politeness and uncertainty when issuing a command.

Feature	ASD Average Rating	TD Average Rating
<b>Request for Information (196 utterances)</b>		
Politeness	<b>2.05</b>	2.04
Uncertainty	<b>2.12</b>	2.1
Information Content	2.01	<b>2.03</b>
<b>Providing Information (335 utterances)</b>		
Politeness*	<b>2.01</b>	1.99
Uncertainty	<b>1.37</b>	1.33
Information Content***	2.0	<b>2.26</b>
<b>Command (127 utterances)</b>		
Politeness***	<b>1.73</b>	1.40
Uncertainty**	<b>1.46</b>	1.29
Information Content	2.02	<b>2.1</b>

Table 5: Results of significance testing for manually annotated linguistics features in individual speech acts. One asterisk indicates a significant difference between the two groups ( $p < 0.05$ ). Two asterisks (\*\*) indicates a high significant difference between the two groups ( $p < 0.01$ ).

## 5. Conclusion

Our results indicate that both manual and automated analysis of conversational data in a collaborative environment can reveal interesting and telling differences between the language use of high-functioning adults with autism spectrum disorder and their matched neurotypical peers. These findings provide the beginnings of quantitative support for the qualitative observations that are routinely made in clinical settings. By being able to identify atypical linguistic characteristics of specific utterances in a collaborative work scenario, our methods can contribute to the development of tools for remediating weaknesses in communication for adults with ASD, a historically underserved population. The data that we will release, including both the transcripts and the manual annotations, will serve as a resource for other researchers working to better understand the communicative challenges facing adults with ASD as they seek to find employment and live independently.

## 6. Acknowledgements

We thank Jill Aldrich for her work recruiting and running participants; and Erin Corcoran, Emily Fabius, and Hannah Jiang for their transcription and annotation of the audio recordings. The project is supported in part by the National Institute on Deafness and Other Communication Disorders (NIDCD) of the National Institutes of Health under award number R21DC017000. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

## 7. References

- Adams, C. (2002). Practitioner review: The assessment of language pragmatics. *Journal of child psychology and psychiatry*, 43(8):973–987.
- Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G. M., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S., and Weinert, R. (1991). The hcrct map task corpus. *Language and Speech*, 34(4):351–366.
- Bishop, D. V., Chan, J., Adams, C., Hartley, J., and Weir, F. (2000). Conversational responsiveness in specific language impairment: Evidence of disproportionate pragmatic difficulties in a subset of children. *Development and psychopathology*, 12(2):177–199.
- Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glott. Int.*, 5(9):341–345.
- Canfield, A. R., Eigsti, I.-M., de Marchena, A., and Fein, D. (2016). Story goodness in adolescents with autism spectrum disorder (asd) and in optimal outcomes from asd. *Journal of Speech, Language, and Hearing Research*, 59(3):533–545.
- Cohen, J. (1968). Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.
- Eales, M. J. (1993). Pragmatic impairments in adults with childhood diagnoses of autism or developmental receptive language disorder. *Journal of autism and developmental disorders*, 23(4):593–617.
- Goodkind, A., Lee, M., Martin, G. E., Losh, M., and Bicknell, K. (2018). Detecting language impairments in autism: A computational analysis of semi-structured conversations with vector semantics. *Proceedings of the Society for Computation in Linguistics*, 1(1):12–22.
- Gorman, K., Olson, L., Hill, A. P., Lunsford, R., Heeman, P. A., and van Santen, J. P. (2016). Uh and um in children with autism spectrum disorders or language impairment. *Autism Research*, 9(8):854–865.
- Hendricks, D. (2010). Employment and adults with autism spectrum disorders: Challenges and strategies for success. *Journal of Vocational Rehabilitation*, 32(2):125–134.
- Kanner, L. (1943). Autistic disturbances of affective content. *Nervous Child*, 2:217–250.
- Ketelaars, M. P., Cuperus, J., Jansonius, K., and Verhoeven, L. (2010). Pragmatic language impairment and associated behavioural problems. *International Journal of Language & Communication Disorders*, 45(2):204–214.
- Landa, R. (2000). Social language use in asperger syndrome and high-functioning autism. *Asperger syndrome*, pages 125–155.
- Lord, C. and Paul, R. (1997). Language and communication in autism. In D. Cohen et al., editors, *Handbook of autism and pervasive developmental disorders*, pages 195–225. Wiley.
- Lord, C., Rutter, M., DiLavore, P., and Risi, S. (2002). *Autism Diagnostic Observation Schedule (ADOS)*. Western Psychological Services.
- Losh, M. and Gordon, P. C. (2014). Quantifying narrative ability in autism spectrum disorder: A computational linguistic analysis of narrative coherence. *Journal of autism and developmental disorders*, 44(12):3016–3025.
- Meyers, B. S., Munaiah, N., Prud’hommeaux, E., Meneely, A., Wolff, J., Alm, C. O., and Murukannaiah, P. (2018). A dataset for identifying actionable feedback in collaborative software development. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 126–131.
- Mok, P. L., Pickles, A., Durkin, K., and Conti-Ramsden, G. (2014). Longitudinal trajectories of peer relations in children with specific language impairment. *Journal of Child Psychology and Psychiatry*, 55(5):516–527.
- Parish-Morris, J., Liberman, M., Ryant, N., Cieri, C., Bateman, L., Ferguson, E., and Schultz, R. T. (2016). Exploring autism spectrum disorders using hlt. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 74–84.
- Prud’hommeaux, E., Morley, E., Rouhizadeh, M., Silverman, L., van Santeny, J., Roarkz, B., Sproatz, R., Kauper, S., and DeLaHunta, R. (2014). Computational analysis of trajectories of linguistic development in autism. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 266–271. IEEE.
- Rutter, M., Bailey, A., and Lord, C. (2003). *Social Communication Questionnaire (SCQ)*. Western Psychological Services, Los Angeles.
- Simmons, E. S., Paul, R., and Volkmar, F. (2014). Assessing pragmatic language in autism spectrum disorder: The yale in vivo pragmatic protocol. *Journal of Speech, Language, and Hearing Research*, 57(6):2162–2173.
- Vincze, V. (2014). Uncertainty detection in hungarian texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1844–1853.
- Volden, J. and Lord, C. (1991). Neologisms and idiosyncratic language in autistic speakers. *Journal of Autism and Developmental Disorders*, 21:109–130.
- Whitehouse, A. J., Watt, H. J., Line, E., and Bishop, D. V. (2009). Adult psychosocial outcomes of children with specific language impairment, pragmatic language impairment and autism. *International Journal of Language & Communication Disorders*, 44(4):511–528.
- Young, E., Diehl, J., Morris, D., Hyman, S., and Bennetto, L. (2005). Pragmatic language disorders in children with autism: The use of two formal tests to distinguish affected children from controls. *Language, Speech, and Hearing Services in Schools*, 36:62–72.

# Automatic Classification of Primary Progressive Aphasia Patients Using Lexical and Acoustic Features

Sunghye Cho<sup>1</sup>, Naomi Nevler<sup>2</sup>, Sanjana Shellikeri<sup>2</sup>, Sharon Ash<sup>2</sup>, Mark Liberman<sup>1</sup>, Murray Grossman<sup>2</sup>

Linguistic Data Consortium<sup>1</sup>, Penn Frontotemporal Degeneration Center<sup>2</sup>  
University of Pennsylvania, 3400 Walnut Street, Philadelphia, PA, USA

{csunghye, myl}@ldc.upenn.edu

{naomine, Sanjana.Shellikeri, ash, mgrossma}@penndmedicine.upenn.edu

## Abstract

Two variants of primary progressive aphasia (PPA) are subtypes of frontotemporal degeneration (FTD), which is the most common type of dementia among individuals under 60 years of age. Semantic variant PPA (svPPA) patients present with semantic deficits in single word use, whereas nonfluent/agrammatic PPA (naPPA) patients produce simplified speech with frequent speech errors and slow speech rates. In this study, we built machine learning systems to classify PPA patients (n=63) and healthy elderly controls (n=36). We automatically extracted 18 lexical and 21 acoustic features with a natural language processing library and a speech activity detector, and we trained classifiers, experimenting with various feature selection and reduction techniques. Our models showed high accuracy, correctly distinguishing patients from controls in more than 90% of cases, svPPA patients from naPPAs in about 89% of cases, and controls, svPPA, and naPPA patients in more than 80% of cases. Our results show that classification of PPA patients using automatically derived linguistic features from digitized speech samples is highly promising, and could potentially be applied in community settings for prescreening. We plan to extend this project by including more features and additional FTD subgroups in the near future.

**Keywords:** Primary progressive aphasia, automatic classification, narrative speech

## 1. Introduction

Frontotemporal degeneration (FTD) is a type of focal dementia caused by atrophy in the brains' frontal and temporal lobes. It is the most common type of neurodegenerative disease among people under 60 years of age (Ratnavalli et al., 2002). Since individuals diagnosed with FTD are relatively young, usually still in the workforce, the personal and societal costs of the disease are substantial. For example, FTD diagnosis often results in early departure from the workforce, increasing economic burden for a household with an FTD patient and negatively affecting not only patients but also the quality of life of their families (Galvin et al., 2017). Because there are no disease-modifying drugs approved for FTD, earlier screening and slowing the apparent disease progression rate through behavioral adjustments to the environment are key to helping patients and their families. This paper proposes three machine learning systems to automatically classify two subgroups of FTD that could potentially be applied in prescreening.

About half of patients with FTD present with a linguistic impairment known as primary progressive aphasia (PPA), and sometimes this can be accompanied by a social-behavioral impairment known as behavioral variant FTD (bvFTD). There are several variants of PPA. Among these subgroups, semantic variant PPA (svPPA) patients are characterized by impaired confrontation naming, frequent substitution of pronouns for nouns, and difficulty in processing concrete words, although they show intact prosody and syntax (e.g., Amici et al., 2007; Bonner et al., 2016; Cousins et al., 2016; Nevler et al., 2019). Nonfluent/agrammatic PPA (naPPA) patients, on the other hand, present with effortful speech, slow speech rates, frequent speech errors, simplified grammar, and difficulty in retrieving verbs (e.g., Ash et al., 2009; Grossman et al., 1996; Rhee et al., 2001). Patients with either of the two subtypes have frontotemporal lobar

degeneration spectrum pathology, which is commonly associated with misfolding of TDP-43 or tau proteins.

Since PPA patients show salient linguistic characteristics, we would expect automatic classification by means of linguistic features to yield high levels of accuracy. There are a few previous studies that have pursued this approach. Fraser et al. (2014) extracted 58 lexical and semantic features from the speech samples of 10 svPPA and 14 naPPA patients and 16 controls. The authors trained classifiers only with significant features for three different tasks: control versus svPPA, control versus naPPA, and svPPA versus naPPA. Their models for controls versus svPPA/naPPA showed high levels of accuracy, from 90% to 100%. However, their best performance for classifying svPPA and naPPA patients was only 79.2% accurate, suggesting that classifying patient groups is more difficult than distinguishing patients from controls. Peintner et al. (2008) extracted 41 acoustic, 81 LIWC (Language Inquiry and Word Count; Pennebaker et al., 2001), and 11 lexical features from 39 participants (9 bvFTD, 8 naPPA, 13 svPPA, and 9 controls), and trained classifiers for various classification tasks. Their composite feature set (significant features from each feature set) showed accuracy over 90% in most classification tasks, except control versus bvFTD and four-way classification. However, they did not list what features were used in the composite set, making it difficult to reproduce their results. Themistocleous et al. (2019) extracted 14 acoustic features, such as mean fundamental frequency and amplitude differences between the first and second harmonics, from 50 patients (17 logopenic variant PPA (lvPPA), 14 svPPA, 11 naPPA, and 8 naPPA with apraxia of speech) and trained classifiers with 3-fold group cross validations and a one-against-all strategy. Their models correctly identified naPPA 82% of the time and svPPA 66% of the time. The authors only used acoustic features, which explains why the accuracy of svPPA patients, who rarely show impairments in prosody, was relatively low. More importantly, all previous studies have had relatively small datasets, raising

the question of whether their results could be generalized to larger datasets. In this paper, we studied 99 participants (63 patients and 36 controls) to investigate whether lexical and acoustic features could predict the diagnostic status of the participants.

## 2. Objectives

Our objectives were to train three predictive models for classifying (1) controls vs. patients, (2) svPPA vs. naPPA patients, and (3) controls, svPPA and naPPA patients, experimenting with different feature selection and reduction techniques, and to identify predictive features for classifying PPA patients.

## 3. Methods

### 3.1 Participants

Our participants consisted of 63 patients diagnosed clinically with either svPPA or naPPA and 36 healthy elderly controls. Forty-two of the 63 patients had svPPA and 21 were naPPA patients. The patients were diagnosed by experienced neurologists at the Department of Neurology of the Hospital of the University of Pennsylvania in accordance with published criteria (Gorno-Tempini et al., 2011). Of the 42 svPPA patients, 32 showed concomitant mild behavioral symptoms, which is a common co-occurrence in this group. We focused on frontotemporal lobar degeneration (FTLD) spectrum pathology in this study, and so we did not include lvPPA patients, who most often have Alzheimer’s pathology. Our participants were matched on sex ratio and education levels, but not on age, because naPPA patients on average have an later disease onset than svPPA patients (Johnson et al., 2005). The patient groups did not differ on the Mini Mental State Exam scores (MMSE) or disease durations, but they significantly differed on the Boston Naming Test (BNT) scores, which is expected due to svPPA patients’ difficulty in naming tasks. All participants were native speakers of English. The study was approved by the Institutional Review Board of the Hospital of the University of Pennsylvania, and all participants signed a written consent form. Participants’ demographic and neuropsychological characteristics are summarized in Table 1.

	controls	svPPA	naPPA	<i>p</i> -value
Age	68.5 (7.9)	63.3 (6.9)	70.4 (9.4)	0.001
Sex	23 F/13 M	23 F/19 M	11 F/11 M	0.483
Education (years)	15.9 (2.5)	15.1 (2.8)	15.3 (3.1)	0.408
MMSE (range: 0-30)	29.2 (1)	22.1 (6.3)	22.7 (5.9)	<0.001
BNT (range: 0-30)	27.9 (2.5)	7.5 (6.4)	24.7 (4.6)	<0.001
Disease duration (yrs)	NA	3.9 (2)	3.2 (1.9)	0.214
Total number of words in Cookie Theft	174.4 (66.4)	148.1 (62.8)	91.0 (55.8)	<0.001

Table 1: Mean (SD) demographic and neuropsychological characteristics of the participants. MMSE: Mini Mental State Exam, BNT: Boston Naming Test.

### 3.2 Data

The Cookie Theft picture from the Boston Diagnostic Aphasia Examination (Goodglass et al., 1983) was used to elicit narrative speech from the participants. Participants described the picture for about one minute, and their descriptions were digitally recorded. Some patients made several recordings, but we used the earliest recording of each participant in this analysis in order to differentiate among the conditions early in the disease course. An expert linguist generated verbatim transcription of the picture descriptions, including all non-verbal speech, hesitations and false starts, and a team of trained annotators at the Linguistic Data Consortium (LDC) of the University of Pennsylvania reviewed and revised the annotations for quality checking.

## 4. Feature Extraction

### 4.1 Lexical Features

We ran a POS tagger in spaCy (Honnibal & Johnson, 2015) to automatically tag POS categories of all words that the participants produced in the picture descriptions. Before running the tagger, we cleaned the transcripts by removing interviewers’ prompts and annotations for non-verbal speech. A professional linguist manually validated the accuracy of spaCy with a subset of our data (n=21). The mean group accuracy varied from 95% (controls) to 90% (PPAs). There was no significant difference in the accuracy among patient groups ( $p>0.05$ ). Since the accuracy of the spaCy POS tagger with their basic model (`en_core_web_sm`) was high, we did not train a POS tagger separately in this study. The POS tokens were tallied per participant, and the count of each POS category per 100 words was calculated ( $= (\text{raw counts}/\text{total number of words}) * 100$ ). In addition to the frequency of each POS category, we measured the number of tense-inflected verbs and unique nouns per 100 words. We summed the number of modal auxiliary verbs, past tense verbs and present tense verbs that spaCy tagged to count the number of tense-inflected verbs per 100 words. The number of noun lemmas was used for the number of unique nouns per 100 words.

We also rated nouns that participants produced for concreteness (Brysbaert et al., 2014), semantic ambiguity (Hoffman et al., 2013), word frequency (Brysbaert & New, 2009), age of acquisition (AoA; Brysbaert et al., 2018) and word familiarity (Brysbaert et al., 2018) for their potential to distinguish svPPA patients from others. Since the published norms we used had a limited number of words, we rated the lemma of a noun if a noun itself was not listed in the published norms. A noun was not rated if neither the noun nor its lemma was listed in the norms. In total, we had 18 text-related features: POS counts per 100 words (nouns, verbs, adjectives, adverbs, prepositions, determiners, conjunctions, interjections, pronouns, and speech errors/partial words—[X] in spaCy), number of tense-inflected verbs and unique nouns per 100 words, lexical features of nouns (concreteness, ambiguity, frequency, AoA, familiarity), and total number of words.

## 4.2 Acoustic Features

We used an in-house Gaussian Mixture Models-Hidden Markov Models based Speech Activity Detector (SAD) developed at the LDC to segment the recordings into speech and silent pause segments. We set the minimum duration of a speech segment at 250 ms and that of a silence segment at 150 ms. We reviewed the outputs of SAD, corrected wrong segmentations, and excluded interviewers’ speech and non-verbal speech segments. Using the durations of speech and silent pause segments, we extracted 12 durational features:

- The mean duration of speech and pause segments
- The number of total pauses and speech segments
- Total speech time (speech only)
- Total pause time (pause only)
- Total time (speech time + pause time)
- Sample duration (duration of the entire recording)
- Percent of speech time of the total time
- Breath frequency (= number of pauses over total time)
- Speech frequency (= number of speech segments over total time)
- Pause rate per minute (= number of pauses over total speech time)

We also pitch-tracked speech segments of the participants with a script in Praat (Boersma & Weenink, 2020) and extracted the 10<sup>th</sup> to 90<sup>th</sup> fundamental frequency (f0) percentile values for each speaker. To minimize individual differences in pitch due to physiological factors, such as sex, height, and the size of the larynx, the extracted f0 values in Hz were converted to semitones (St) using each speaker’s 10<sup>th</sup> percentile as a baseline:  $St = \log_2(\text{Hz} / 10^{\text{th}} \text{percentile}) * 12$ . We had 21 acoustic features in total, including pitch percentile values along with the 12 durational features. The final feature set included 18 lexical and 21 acoustic features and 3 demographic characteristics of the participants: age, sex, and education level.

## 5. Model Training

We trained two different machine learning algorithms from the scikit-learn package (Pedregosa et al., 2011) in Python: Random Forest and Support Vector Machine (SVM) classifiers. In all models, we imputed missing values with a mean value using SimpleImputer and standardized features with StandardScaler in scikit-learn for effective learning. We performed leave-one-out cross-validation (CV) to evaluate the generalizability of the models and reported the average accuracy of all CV folds.

We experimented with feature selection and reduction methods. For feature selection, we performed t-tests (for binary classifications) and trained models with features that were significant at the level of  $p < 0.05$ , 0.01, 0.005, and 0.001. We used the same feature set used in the control-patient pairwise classification for the three-way classification (control vs. svPPA vs. naPPA). We compared the performance of models trained with selected features and a model without any feature selection. For feature reduction, we performed Principal Component Analysis (PCA) and trained models, varying the number of components from 1 to 10. We compared the performance of models trained with PCA components and that of a

model trained without any feature reduction and reported the best performance after tuning hyperparameters.

## 6. Classification Results

### 6.1 Binary Classification between Controls and Patients

An SVM classifier trained with all features which were reduced to 10 PCA components performed best in this classification task, showing 90.9% accuracy and 0.94 area under the curve (AUC). Our model correctly identified 33 controls out of 36 and 57 patients out of 63. The full classification report is shown in Table 2, and the receiver operating characteristic (ROC) curve for this contrast is provided in Figure 1.

	Accuracy	Precision	Recall	F1-score
Controls	0.92	0.85	0.92	0.88
Patients	0.90	0.95	0.90	0.93
Weighted average	0.91	0.91	0.91	0.91

Table 2: Classification report of the SVM classifier for the classification of patients and controls.

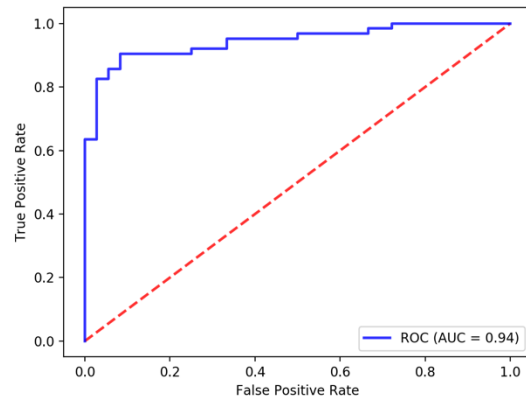


Figure 1: Receiver Operator Characteristic Curve for the classification of controls and patients.

### 6.2 Binary Classification of Patient groups

A Random Forest classifier trained with features that were significant at the level of  $p < 0.005$  and reduced to three PCA components performed best in this classification task. The model showed 88.9% accuracy with 0.87 AUC. The model correctly identified 40 svPPA patients out of 42 and 16 naPPA patients out of 21. Our model resulted in a higher F1-score for classifying svPPA patients (0.92) than naPPA patients (0.82), suggesting that in general identifying naPPA patients was more difficult than identifying svPPA patients. The full classification scores are in Table 3, and the ROC curve for this contrast is provided in Figure 2.

	Accuracy	Precision	Recall	F1-score
svPPA	0.95	0.89	0.95	0.92
naPPA	0.76	0.89	0.76	0.82
Weighted average	0.89	0.89	0.89	0.89

Table 3: Classification report of the Random Forest classifier for the classification of svPPA and naPPA patients.

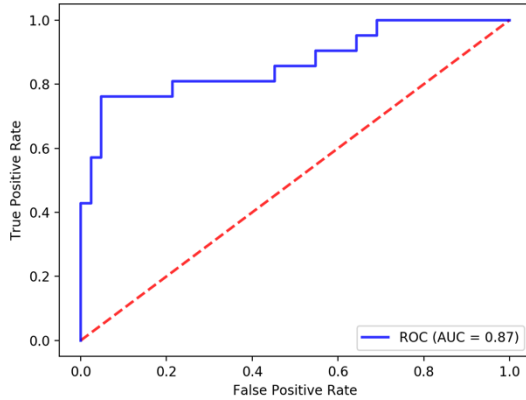


Figure 2: Receiver Operator Characteristic Curve for the classification of svPPA and naPPA patients.

The features that were selected included counts of nouns, pronouns, verbs, tense-inflected verbs, speech errors/partial words, unique nouns per 100 words; concreteness, semantic ambiguity, frequency of nouns; participants' age and total number of pauses. Figure 3 shows group differences in the selected features.

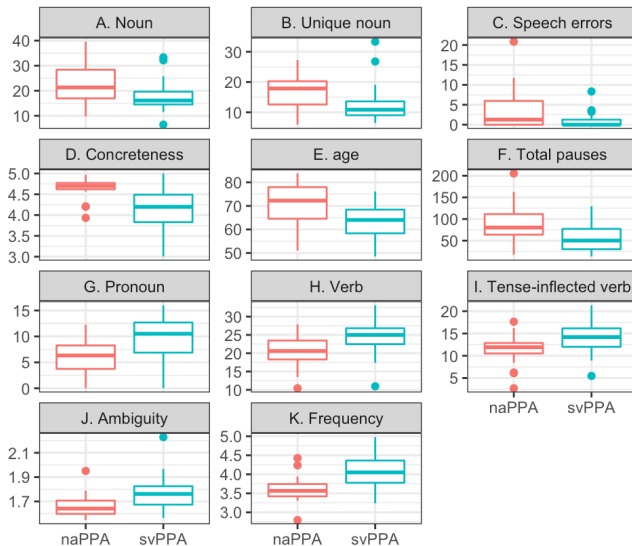


Figure 3: Group differences in selected features for the classification of svPPA and naPPA patients. The POS counts and the numbers of tense-inflected verbs and unique nouns are per 100 words. The top two rows show features where values of naPPA patients are significantly higher than those of svPPA and the bottom two rows show features where values of svPPA patients are

significantly higher than those of naPPA (both at  $p < 0.005$ ).

Among the 11 selected features, most were lexical, and only one acoustic feature, total number of pauses, was selected. As expected, semantic aspects of nouns that patients produced, such as concreteness and semantic ambiguity, were important features in distinguishing svPPA patients from naPPA patients. Further discussion of the acoustic features in PPA patients can be found in Nevler et al. (2019), and further discussion of the lexical features can be found in Cho et al. (under review).

### 6.3 Three-way Classification

An SVM classifier trained with all features without any feature reduction performed best for the three-way classification, yielding 80.8% accuracy with 0.9 macro-averaged AUC. The model correctly identified 32 controls out of 36, 34 svPPA patients out of 42, and 14 naPPA patients out of 21. The model's F1-score is high for controls and svPPA patients ( $> 0.8$ ), but it was below 0.7 for naPPA patients, again suggesting that naPPA patients were difficult to identify. The full classification report and the confusion matrix are provided in Tables 4 and 5, and the ROC curve for this contrast is provided in Figure 4.

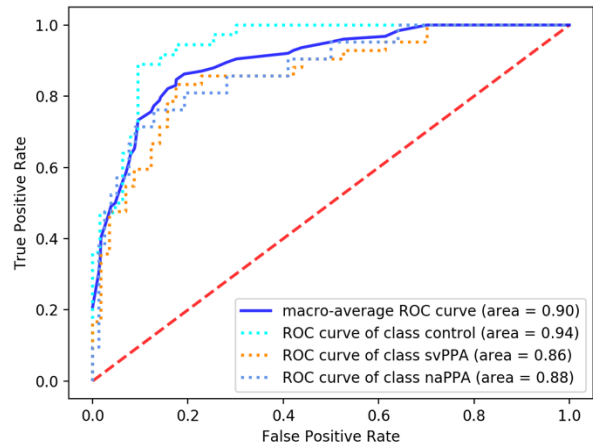


Figure 4: Receiver Operator Characteristic Curve for the classification of controls and svPPA and naPPA patients.

	Accuracy	Precision	Recall	F1-score
Control	0.89	0.84	0.89	0.86
svPPA	0.81	0.83	0.81	0.82
naPPA	0.67	0.70	0.67	0.68
Weighted average	0.81	0.81	0.81	0.81

Table 4: Classification report of the SVM classifier for the three-way classification.

	Control	svPPA	naPPA
Controls	32	2	2
svPPA	4	34	4
naPPA	2	5	14

Table 5: Confusion matrix of the three-way classification (column: actual, row: predicted). The number of accurately classified participants is highlighted in gray.

## 7. Discussion and Conclusion

This paper reports the results of automatic classification systems for three classification tasks: i) control versus patients, ii) svPPA versus naPPA patients, and iii) control versus svPPA versus naPPA. We automatically extracted 18 lexical features from one-minute narrative speech samples using spaCy, one of the most modern, state-of-the-art natural language processing libraries in Python. We also automatically extracted 21 acoustic and durational features with SAD. Using these features with additional demographic information, we trained three machine learning classifiers, experimenting with different feature selection and reduction techniques, and used leave-one-out cross-validation. We found group differences in the selected features. Our model for the control versus patient classification trained with all features, which were reduced to 10 PCA components, correctly distinguished patients from controls in more than 90% of cases. Our classifier for the svPPA versus naPPA task selected 11 features (9 lexical, 1 acoustic and 1 demographic), which were later reduced to 3 PCA components. Our classifier correctly identified the diagnostic group of the patients with 88.9% accuracy, which outperformed the system for the same task in previous studies (79.2% in Fraser et al., 2014; 82% for naPPA patients in Themistocleous et al., 2019). Lastly, our system for the three-way classification, which was trained with all features without any feature reduction, showed high overall accuracy (over 80%) in classifying controls, svPPA and naPPA patients, which is much higher than the chance level (33.3%). The performance of the systems in this report is highly promising in that we only had one-minute narrative speech samples, which are quick and easy to collect. We believe that this line of research could potentially benefit populations with the earliest features of PPA.

Our models performed well, but there is still room for improvement, in particular, for the three-way classification system, where classification of naPPA was < 80%. In the future, we plan to include more features, such as letter or category fluency scores, Mel-frequency cepstral coefficients, or word frequency as log-odds ratio (Monroe et al., 2008) to improve the performance of the models. We also aim to extend our research by including more patient groups. First, we would consider evaluating patients with lvPPA, which is another variant of PPA associated with Alzheimer’s disease pathology, with frequent filler words (*um* or *uh*) as a prominent feature. Second, we would consider including bvFTD patients, who have pathology

similar to that of svPPA and naPPA patients. Although without obvious aphasia, these patients do have subtle speech deficits (Nevler et al., 2018). In addition, we plan to collect conversational data in the near future to explore subtle group differences among these patient groups that may not have been captured in monologue, narrative speech samples. In natural conversation, speakers employ a variety of prosodic features to deliver the intended message effectively. We believe these additional features will improve the models’ performance.

## 8. Acknowledgements

We thank the participants and their family members for participating in the study and the research assistants who helped collect the data. This study was funded by National Institutes of Health (AG017586, AG053940, AG052943, NS088341, DC013063, AG054519), the Institute on Aging at the University of Pennsylvania, the Alzheimer’s Association (AACSF-18-567131), an anonymous donor, and the Wyncote Foundation.

## 9. Bibliographical References

- Amici, S., Ogar, J., Bozeat, S., Arnold, R., Watson, P., and Hodges, J.R. (2006). Performance in specific language tasks correlates with regional volume changes in progressive aphasia. *Cognitive and Behavioral Neurology*, 20(4): 203–211.
- Ash, S., Moore, P., Vesely, L., Gunawardena, D., McMillan, C., Anderson, C., ... and Grossman, M. Non-fluent speech in frontotemporal lobar degeneration. *Journal of Neurolinguistics*, 22(4):370–383.
- Boersma, P., and Weenink, D. (2020). Praat: doing phonetics by computer [Computer program]. Version 6.1.09, retrieved 26 January 2020 from <http://www.praat.org>
- Bonner, M., Price, A., Peelle, J., and Grossman, M. (2016). Semantics of the visual environment encoded in parahippocampal cortex. *Journal of Cognitive Neuroscience*, 28(3): 361–378.
- Brysbaert, M., and New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4):977–990.
- Brysbaert, M., Mandera, P., and Keuleers, E. (2018). Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods*, 51(2):467–479.
- Brysbaert, M., Warriner, A. B., and Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3):904–911.
- Cho, S., Nevler, N., Ash, S., Shellikeri, S., Irwin, D., Massimo, L., Rascovsky, K., Olm, C., Grossman, M., and Liberman, M. (under review). Automated analysis of lexical features in Frontotemporal Degeneration.
- Cousins, K., York, C., Bauer, L., and Grossman, M. (2016). Cognitive and anatomic double dissociation in the representation of concrete and abstract words in semantic variant and behavioral variant frontotemporal degeneration. *Neuropsychologia*, 84:244–251.
- Fraser, K., Meltzer, J., Graham, H., Leonard, C., Hirst, G., Black, S., and Rochon, E. (2014). Automated

- classification of primary progressive aphasia subtypes from narrative speech transcripts. *Cortex*, 55:43–60.
- Galvin, J., Howard, D., Denny, S., Dickinson, S., and Tatton, N. (2017). The social and economic burden of frontotemporal degeneration. *Neurology*, 89:2049–2056.
- Goodglass, H., Kaplan, E., and Weintraub, S. (1983). Boston diagnostic aphasia examination. Philadelphia, PA, Lea & Febiger.
- Gorno-Tempini, M. L., Hillis, A. E., Weintraub, S., Kertesz, A., Mendez, M., Cappa, S. F., et al. (2011). Classification of primary progressive aphasia and its variants. *Neurology*, 76:1006–1014.
- Grossman, M., Mickanin, J., Onishi, K., Hughes, E., D’Esposito, M., Ding, X. S., ... and Reivich, M. (1996). Progressive nonfluent aphasia: Language, cognitive, and PET measures contrasted with probable Alzheimer’s disease. *Journal of Cognitive Neuroscience*, 8(2) :135–154.
- Honnibal, M., and Johnson, M. (2015). An improved non-monotonic transition system for dependency parsing. In Mårques et al., Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1373–1378, Lisbon, Portugal, September.
- Johnson, J.K., Diehl, J., Mendez, M.F., Neuhaus, J., Shapira, J.S., Forman, M., Chute, D.J., Roberson, E.D., Pace-Savitsky, C., Neumann, M., Chow T.W., Rosen, H.J., Forstl, H., Kurz, A., and Miller, B.L. Frontotemporal lobar degeneration: Demographic characteristics of 353 patients. *Arch Neurol.*, 62: 925–930.
- Monroe, B., Colaresi, M., and Quinn, K. (2008). Fightin’ words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis* 16:372–403.
- Nevler, N., Ash, S., Irwin, D. J., Liberman, M., and Grossman, M. (2019). Validated automatic speech biomarkers in primary progressive aphasia. *Annals of Clinical and Translational Neurology*, 6(1):4–14.
- Nevler, N., Ash, S., Jester, C., Irwin D. J., Liberman, M., and Grossman, M. (2018). Automatic measurement of prosody in behavioral variant FTD. *Neurology*, 89:650–656.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, B., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011) Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Peintner, B., Jarrold, W., Vergyri, D., Richey, C., Gorno-Tempini, M., and Ogar, J. (2008). Learning diagnostic models using speech and language measures. In Wheeler B., et al., Proceedings of the 30<sup>th</sup> Annual International IEEE EMBS Conference, pages 4648–4651, Vancouver, Canada, August.
- Ratnavalli, E., Brayne, C., Dawson, K., and Hodges, J.R. (2002). The prevalence of frontotemporal dementia. *Neurology*, 58:1615–1621.
- Rhee, J., Antiquena, P., and Grossman, M. (2001). Verb comprehension in frontotemporal degeneration: The role of grammatical, semantic and executive components. *Neurocase*, 7(2):173–184.
- Themistocleous, C., Ficek, B., Webster K.T., Wendt H., Hillis A.E., Den Ouden, D.B., and Tsapkini, K. Acoustic markers of PPA variants using machine learning.

# Affective Speech for Alzheimer’s Dementia Recognition

Fasih Haider, Sofia de la Fuente, Pierre Albert, Saturnino Luz

Usher Institute, Edinburgh Medical School

the University of Edinburgh

{Fasih.Haider, sofia.delafuente, pierre.albert, S.Luz}@ed.ac.uk

## Abstract

Affective behaviour could provide an indicator of Alzheimer’s disease and help develop clinical tools for automatically detecting and monitoring disease progression. In this paper, we present a study of the predictive value of emotional behaviour features automatically extracted from spontaneous speech using an affect recognition system for Alzheimer’s dementia detection. The effectiveness of affective behaviour features for Alzheimer’s Disease detection was assessed on a gender and age balanced subset of the *Pitt Corpus*, a spontaneous speech database from DementiaBank. The affect recognition system was trained using the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) and the Berlin database of emotional speech. The output of this system provides classification scores or class posterior probabilities of 6+1 emotions as an input for statistical analysis and Alzheimer’s dementia detection. The statistical analysis shows that the non-AD subjects have higher mean value of classification scores for *anger* and *disgust*, along with a higher entropy of classification scores than AD subjects. The AD subjects have a higher classification scores for the *sad* emotional behaviour than non-AD. This paper also introduces a novel ‘affective behaviour representation’ feature vector for Alzheimer’s dementia recognition. Results show that classification models based solely on affective behaviour attain 63.42% detection accuracy.

**Keywords:** Affective Computing, Social Signal Processing, Dementia, Alzheimer, Cognitive Decline Detection, Cognitive Impairment Detection

## 1. Introduction

Dementia is a category of neurodegenerative diseases characterised by long-term and usually gradual decrease of cognitive functioning (American Psychiatric Association, 2000). Whilst memory loss is frequently considered the most prominent symptom of dementia, in particular of Dementia of the Alzheimer Type (DAT), speech and language alterations are also common (Kirshner, 2012). For instance, word-finding difficulties (i.e. anomia) are reported from early stages of cognitive impairment, when patients describe how they can see certain words “floating in front of them”, although they do not manage to “catch” them in order to put them in a sentence. Literature also suggests that patients with DAT have difficulty accessing semantic information when they intend to do so (Bondi et al., 1996). Since successful communication is essential for meaningful social interaction, this takes a toll on patients’ and their carers’ wellbeing.

It is presumed that such difficulties increase the level of frustration and have an impact on the emotional life of these patients. The prevalence of apathy, dysphoria and depression in Alzheimer’s Disease (AD) increases with the severity of the condition (Landes et al., 2005). In fact, the overlap between apathy and depression becomes particularly prominent in this clinical population (Starkstein et al., 2005). These comorbidities are relatively well established and have spurred research on differential diagnosis between dementia and depression (Leyhe et al., 2017).

One of the reasons suggested in an attempt to explain the comorbidity between dementia and depression is the role of emotions in memory encoding, since both conditions progress with an increased forgetfulness (Hart et al., 1987). Emotional abilities, amongst which are the expression of our own emotions as well as the recognition of those of others, are also decisive for social communication (Lopes

et al., 2004). Emotional information can be conveyed in different ways, from explicit facial and verbal expression (e.g. smile, pout, happy statement) to more subtle non-verbal cues, such as intonation, modulation of vocal pitch and loudness of emotional expression. These non-verbal cues are generally referred to as emotional prosody. Both expression and recognition of emotional prosody seem to be impaired in DAT (Horley et al., 2010), though the latter has been more widely studied.

Research on computational speech technology to better characterise emotional prosody could shed a light on the expression of emotions in people with DAT. This study aims to apply a signal processing model for recognition of emotional prosody in DAT with two main objectives. First, we wish to determine whether certain emotions are predominant in DAT whilst others are subdued. With this purpose, we train an emotion recognition model on a high quality dataset of emotion expression, and then use this model to classify speech segments of a dataset containing speech of AD and non-AD participants into 6+1 emotional state labels. Second, once the distribution of emotions across each audio recording is established (i.e. classification scores or class posterior probabilities), this information will be used as an input for a classifier, aimed at automatic detection of DAT based on emotional prosody, as shown in Figure 1.

## 2. Related work

As far as research on emotions is concerned, the most common paradigms tend to rely on facial expression and image processing (e.g. (Seidl et al., 2012)), with less published work on prosody and other linguistic features. However, there is quantifiable evidence that acoustic analysis can give an account of emotional expression. For instance, sadness is associated with lower speech rate and lower mean fundamental frequency ( $F_0$ ) than emotions such as happiness, fear or anger (Juslin and Laukka, 2003).



Figure 1: Pipeline for dementia recognition through Affective/Emotional Behaviour.

Previous research on DAT and emotional prosody has predominantly focused on recognition (*receptive emotional prosody*), as opposed to expression (*expressive emotional prosody*). Findings in both areas have yielded promising but as yet inconclusive results. For instance, research findings pointed at impaired emotional processing in DAT, though still relatively preserved in comparison to cognitive abilities, suggesting that impairment of emotional prosody might be secondary to the decline of another cognitive function (Bucks and Radford, 2004).

Receptive emotional prosody is generally evaluated as the accurate identification of certain emotional tones when someone speaks (Taler et al., 2008). By removing information based on words (i.e. filtering out the spectral energy above a certain frequency), promising results, based solely on prosodic features, have been reported. Not only there are signs of an impaired processing of emotional prosodic information in DAT, but there is also evidence suggesting that such impairment precedes the decline of other linguistic aspects (Testa et al., 2001).

The work presented in this paper focuses on expressive emotional prosody. There are three distinct ways to elicit data: a) prosodic modelling, which requires participants with DAT to repeat a sentence copying a previously heard emotional tone (Testa et al., 2001), b) commanded production, which requires participants to read semantically neutral sentences with a designated emotional tone (Roberts et al., 1996), and c) natural expression, whereby participants are required to describe an emotional experience (Testa et al., 2001). Testa et al. (2001) applied speech analysis to evaluate the quality of emotional expression from participants with DAT, based on prosodic information, and found that receptive prosody was impaired earlier than expressive prosody, but also that both were impaired early in the progression of the disease. However, they used a limited feature set, essentially analysing variability in fundamental frequency.

The same elicitation method, natural expression, was used by a more recent study where participants were asked to share an autobiographical memory (Han et al., 2014). They measure emotional prosody quality of memory retrieval, under the common assumption that traumatic events contain essential information for survival and hence benefit from superior encoding. They report an impaired ability to express emotions in early AD, regardless of whether the autobiographical memory is recent or remote - one of the first studies looking at emotional prosody instead of semantic content of those memories. More importantly, they found a correspondence between emotional expression and cognitive functioning. Another recent work develops a measure for emotional response as part of a comprehensive prosodic account, reporting gradual changes in spontaneous speech

and emotional response as cognition declines (Lopez-de Ipiña et al., 2016). Even though the approach of these more recent studies extends previous research by using multiple acoustic measures, their acoustic feature set is still limited in both size and underlying rationale. While  $F_0$  and its associated measures correlate acoustically to perceived pitch, we propose to use a standardised and theoretically motivated feature set to detect psychological changes in voice production, namely, eGeMAPS (Eyben et al., 2016).

Further research is clearly necessary to provide a solid account of the quality of emotional expression in the context of DAT. A computational approach to this task would lessen the problem of subjectivity and low inter-rater reliability, as well as contributing to a potentially automatic diagnostic support tool. We hypothesise that if the expression of emotions through speech is impaired in a person with AD, a classifier should have greater difficulty distinguishing emotions in the voice of a person with AD than in the voice of a person without AD (non-AD). Therefore, a measure of uncertainty in emotion classification, such as the Shannon entropy of posterior (emotion) class probabilities, might be a suitable feature for a classification model for DAT. Besides, there is controversy about the actual reliability of humans identifying other humans' emotions, with sadness and anger usually being the emotions with highest agreement. Research evidence shows that emotion recognition from voice samples is about 60% accurate (Johnstone and Scherer, 2000), which we will take as a baseline for our model.

### 3. Dataset Description

This section describes the Berlin Database of Emotional Speech, used for the training of our emotion recognition system, and the age and gender balanced subset of the Pitt dataset, used for DAT prediction based on emotional speech features.

#### 3.1. Berlin Database of Emotional Speech (*EmoDB*)

The *EmoDB corpus* (Burkhardt et al., 2005) is a dataset commonly used in the automatic emotion recognition literature. It features 535 acted emotions in German, based on utterances carrying no emotional bias. The corpus was recorded in a controlled environment resulting in high quality recordings, but actors were allowed to move freely around the microphones, affecting absolute signal intensity. In addition to the emotion, each recording was labelled with phonetic transcription using the SAMPA phonetic alphabet, emotional characteristics of voice, segmentation of the syllables, and stress. The quality of the data set was evaluated by perception tests carried out by 20 human participants. In a first recognition test, subjects listened to a recording once before assigning one of the available category, achieving an

average recognition rate of 86%. A second naturalness test was performed. Documents achieving a recognition rate lower than 80% or a naturalness rate lower than 60% were discarded from the main corpus, reducing the corpus to 535 recordings from the original 800. We have normalized all the speech utterances’ volume into the range  $[-1:+1]$  dBFS before acoustic feature extraction. The motivation behind this normalization is to make the model robust against different recording conditions such as distance between the microphone and the subject.

### 3.2. The Pitt Corpus

This study specifically uses the *Pitt Corpus*, gathered longitudinally between 1983 and 1988 on a yearly basis as part of the Alzheimer Research Program at the University of Pittsburgh (Corey Bloom and Fleisher, 2000). Participants are categorised into three groups: dementia, control (non-AD), and unknown status. All participants were required to be above 44 years of age, have at least seven years of education, have no history of nervous system disorders or be taking neuroleptic medication, have an initial MMSE score of 10 or more and be able to provide informed consents. Extensive neuropsychological and physical assessments conducted on the participants are also included; more detailed information of this cohort can be found in (Becker et al., 1994). This study selected only the dementia and control groups for a binary diagnosis of AD and non-AD.

The *Pitt Corpus* contains data elicited through the following tasks: the Cookie Theft stimulus picture description for AD and non-AD groups, and a word fluency task, a story recall task, and a sentence construction task for the AD group only. In this study, we specifically chose the Cookie Theft description task subset. Table 1 lists the data available in this set. Participants were shown the Cookie Theft picture and were asked to describe the picture in their own words.

Table 1: Statistics of the DementiaBank Pitt corpus

	non-AD	AD*
Number of patients	99	194
Number of visits (recordings)	242	307
with 1 visit	26	117
with 2 visits	28	53
with 3 visits	28	12
with 4 visits	9	9
with 5 visits	8	3

\*One participant (ID:172) has changed the diagnosis from "Control" (in the first visit) to "Dementia" (in the remaining 3 visits).

The *Pitt Corpus* includes both the manual transcripts of the clinical sessions and the corresponding audio recordings for both participants (i.e. AD and non-AD) groups. The transcripts comprise both the speech of the Investigator (INV) and the Participant (PAR). Based on the information provided by DementiaBank, the AD and non-AD groups were not matched with age, gender or education. This study will thus create a subset matched for age and gender to eliminate bias.

### 3.3. Subset Selection from Pitt Corpus

The steps taken to select a balanced subset of the *Pitt Corpus*: Cookie Theft task, for our experiment are shown in Figure 2, and described in the remainder of this section.

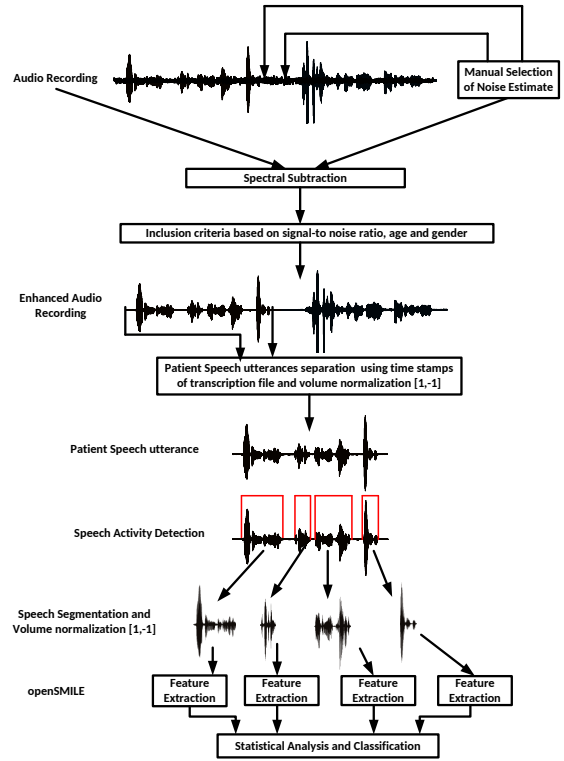


Figure 2: Data pre-processing steps.

#### 3.3.1. Audio Enhancement and inclusion criteria

We manually selected a short interval from each audio recording which contained only the noise and applied spectral subtraction to eliminate that noise. Other non-target sounds such as background talk, ambulance sirens, door slamming, were minimised by selecting audio files with signal-to-noise ratio (SNR) greater than or equal to -17 dB. Where multiple audio files existed per participant only the most recent audio file of that participant was selected.

#### 3.3.2. Matching the Data for Gender and Age

Age and gender are considered major risk factors for dementia (Dukart et al., 2011). Therefore, these variables are possible confounders between the AD and non-AD groups. To eliminate these confounders, we selected a subset of the *Pitt Corpus* in which the AD and non-AD groups are matched for age and gender. Along with the inclusion criteria defined in Section 3.3.1., matching gender and age for both AD and non-AD datasets ensured homogeneity of the sample population, reducing confounding and increasing the likelihood of finding a true association between exposure and outcomes. The age ranges were chosen empirically to optimise the number of recordings included in the final dataset. As a result, 164 participants matched the selection criteria to be included in the study. Of these, 82 were healthy and 82 were diagnosed with probable AD.

After testing the different ranges of the age intervals, the dataset was balanced and could produce the optimal number of recordings by using the age range from 45 to 80 years with the interval of 5 years. Table 2 presents the demographic data. Participants’ age in each group ranged between 50 and 80 years old.

Table 2: Basic characteristics of the patients in each group (AD/non-AD)

Age Interval	AD		non-AD	
	Male	Female	Male	Female
(50, 55)	2	1	2	1
(55, 60)	7	8	7	8
(60, 65)	4	9	4	9
(65, 70)	10	14	10	14
(70, 75)	9	11	9	11
(75, 80)	4	3	4	3
Total	36	46	36	46

### 3.3.3. Speech Segmentation

Speech segmentation was performed on the audio files that met the above described selection criteria. The study only focuses on the participants’ speech; therefore, the investigators’ speech were excluded from further processing. First, we extracted the participants’ speech utterances using the timestamps (start time and end time) from the DementiaBank transcripts. However, as the participants’ speech exhibits long pauses and low volume, we normalised the volume to the range [-1:+1] dBFS and then used speech activity detection (with an energy threshold of 50 dB) <sup>1</sup> for speech segmentation (i.e. to separate speech from pauses). Volume normalization helps tackling different recording conditions, particularly variations in microphone placement in relation to the participant.

### 3.4. Feature Extraction

We used the openSMILE (Eyben et al., 2013) toolkit for the extraction of prosodic features using the eGeMAPS feature set, which is widely used for emotion recognition. The eGeMAPS (Eyben et al., 2016) feature set contains the  $F_0$  semitone, loudness, spectral flux, MFCC, jitter, shimmer,  $F_1$ ,  $F_2$ ,  $F_3$ , alpha ratio, Hammarberg index and slope  $V_0$  features, including many statistical functions applied on these features, which results in a total of 88 features for every speech utterance. We removed features which are correlated ( $|r| > 0.2$ ) with the duration of speech utterances. This left us with 75 remaining acoustic features for further processing.

## 4. Affect Recognition System

The Affect Recognition System was trained using Support Vector Machines (SVM) using the SMO solver with box constraint (k) of 0.75, and linear kernel function. We employed the MATLAB<sup>2</sup> implementation of this classifier, using the statistics and machine learning toolbox. A

<sup>1</sup><https://pypi.org/project/audiotk/> (Last accessed: January 2020)

<sup>2</sup><http://uk.mathworks.com/products/matlab/> (Last accessed: January 2020)

leave-one-subject-out (LOSO) cross-validation procedure was adopted, where the training data do not contain any information on the validation subjects. The results are shown in Figure 3. The affect recognition system provides an accuracy of 69.72% with a Kappa of 0.638.

True Class	Predicted Class							Recall (%)
	Ang.	Bore	Disg.	Fear	Happ.	Sad	Neu.	
Ang.	113		2	6	6			88.98
Bore	1	54	7	4		6	9	66.67
Disg.	4	2	25	7	3	1	4	54.35
Fear	9		3	43	5	1	8	62.32
Happ.	34	1	4	5	24	1	2	33.80
Sad			4	2		54	2	87.10
Neu.	2	8		4	3	2	60	75.95
Precision (%)	69.33	83.08	55.56	60.56	58.54	83.08	70.59	

Accuracy = 69.72%    Kappa = 0.638    UAR = 67.02%

Figure 3: Emotion recognition results.

Once trained on *EmoDB*, our affect recognition system was used to identify emotions in the 4,076 speech segments in our dementia dataset. The results are shown in Figure 4. Noticeably, the AD subjects have more Sad (260 compared to 156) and Happy (616 compared to 580) instances than non-AD, and non-AD subjects have more Anger, Boredom, Disgust and Neutral instances.

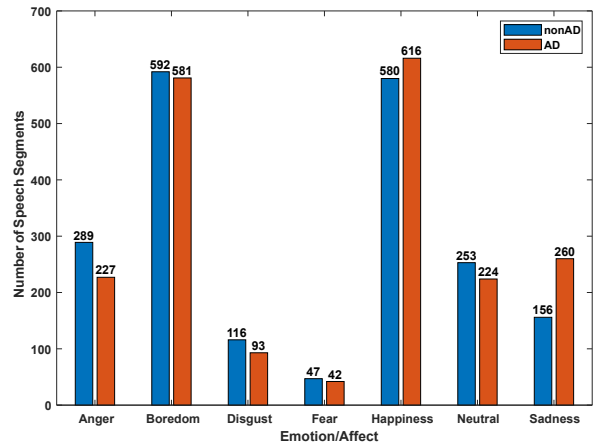


Figure 4: Emotion Recognition on subset of *Pitt Corpus* for AD and non-AD subjects.

## 5. Statistical Analysis

To find the relationship between emotions and AD, we used the matrix of scores (i.e. classification scores or class posterior probabilities) which indicates the likelihood that a speech segment expresses a particular emotion. As a results we have a vector of (1x7) for each speech segment representing likelihood of 6+1 emotions. We also calculated the entropy of the posterior probabilities per speech segment to measure the degree of the model’s ‘uncertainty’ with regards to a classification.

The one-sample Kolmogorov-Smirnov test shows that the data (i.e. classification scores and entropy of scores) follows a normal distribution assumption with  $p < 0.001$ .

Table 3: Statistical Analysis: ANOVA test results; p-adj indicates p values adjusted for multiple comparison by controlling the false discovery rate.

	$H_\phi^{Anger}$	$H_\phi^{Bore}$	$H_\phi^{Disgust}$	$H_\phi^{Fear}$	$H_\phi^{Happy}$	$H_\phi^{Sad}$	$H_\phi^{Neutral}$	$H_\phi^{Entropy}$
nonAD	0.1569	0.2178	0.1174	0.0446	0.1964	0.0815	0.1854	2.4750
AD	0.1285	0.2207	0.1026	0.0446	0.1996	0.1297	0.1742	2.4413
p-value	<b>0.0002</b>	0.7117	<b>0.0050</b>	0.9877	0.5469	< <b>0.0001</b>	0.0642	<b>0.0216</b>
p-adj	<b>0.0011</b>	0.81	<b>0.0133</b>	0.9877	0.729	< <b>0.0001</b>	0.1000	<b>0.0400</b>

We have set the following null hypotheses for the Anova test: the scores of emotion  $x \in \{\text{Anger, Boredom, Disgust, Fear, Happiness, Sadness, Neutral}\}$  do not differ between the AD and non-AD groups ( $H_\phi^x$ ), and the Shannon entropy of the posterior probability distribution for an emotion given a speech segment ( $H_\phi^{Entropy}$ ) does not differ between the AD and non-AD groups. The anova test results are shown in Table 3. The ANOVA test rejects the null hypothesis for Anger, Disgust, Sadness and entropy, when  $p$  values are corrected for multiple comparisons by using the Benjamini-Hochberg procedure to control the false discovery rate. However, no significant differences were found for Boredom, Fear, Happiness and Neutral emotion expressions.

## 6. Affective Behaviour Representation

To aggregate affective behaviour within an audio recording per subject for automatic classification of AD, we propose a novel Affective Behaviour Representation (ABR) feature vector. This consists of the following steps:

1. *Emotion Recognition* of segments: we used an emotion recognition model to recognise emotions within segments using audio features.
2. *Generation* of the Affective Behaviour Number ( $nABR_{A_i}$ ) vector by calculating the number of segments in each emotion category for each audio ( $A_i$ ) i.e. histogram representation of number of speech segments for 6+1 emotions for each audio recording.
3. *Normalisation of segments*: as the number of segments is different for each subject (i.e. the duration of all audio recordings is not constant), we normalise the ( $nABR_{A_i}$ ) by dividing it by the total number of segments present in each audio recording (i.e. the L1 norm of  $nABR_{A_i}$ ), as shown:

$$nABR_{A_i\text{norm}} = \frac{nABR_{A_i}}{\|nABR_{A_i}\|_1} \quad (1)$$

4. *Generation* of the Affective Behaviour Score ( $sABR_{A_i}$ ) vector by summing the score for each emotion category for each audio recording ( $A_i$ ); that is, the histogram representation of scores for 6+1 emotions for each audio recording.
5. *Normalisation of score*: as the number of segments is different for each subject (i.e. the duration of all audio recordings is not constant), we normalise the ( $sABR_{A_i}$ ) by dividing it by the sum of scores

of segments for each audio recording as we did for  $nABR_{A_i}$ :

$$snABR_{A_i\text{norm}} = \frac{sABR_{A_i}}{\|sABR_{A_i}\|_1} \quad (2)$$

6. *Affective Behaviour Representation (ABR)*: we fused the  $nABR_{A_i}$  and  $snABR_{A_i}$  to generate the ABR, as shown in Equation 3

$$ABR_{A_i\text{norm}} = [nABR_{A_i\text{norm}}, snABR_{A_i\text{norm}}] \quad (3)$$

### 6.1. AD Detection

We conducted three classification experiments to detect cognitive impairment due to AD, namely:

1. *Segment Level (SL) classification*: in this experiment we trained and tested our classifiers in a LOSO setting, with scores of emotions to predict whether the speech segments were uttered by a non-AD or AD patient;
2. *Majority Vote (MV) classification*: using the results of segment-level classification, we calculated the number of segments detected as AD and non-AD for each subject and then took a majority vote to assign an overall label to the subject; and
3. *Affective Behaviour Representation*: we generated the ABR using the score and labels of emotion recognition system as described in section 6., and then used  $ABR_{A_i\text{norm}}$  for classification as before.

### 6.2. Classification Methods

The classification experiments were performed using five different methods, namely decision trees (DT, with leaf size of 20), nearest neighbour (KNN with K=1), linear discriminant analysis (LDA), random forests (RF, with 50 trees and a leaf size of 20) and support vector machines (SVM, with a linear kernel with box constraint of 0.1, and sequential minimal optimisation solver). The classification methods were implemented in MATLAB using the statistics and machine learning toolbox. A leave-one-subject-out (LOSO) cross-validation setting was adopted.

### 6.3. Results

The AD recognition results for all three experiments (detailed in Section 6.1.) are shown in Table 4. It is noted that the ABR (59.76) provides better results than MV (58.54) and SL (52.70). The random forest classifier provides the best results for all three experiment. We have selected top three classifiers (57.93% for MV using KNN, 58.54% for

True Class		Predicted Class		Recall
		nonAD	AD	
nonAD	nonAD	60	22	73.17%
	AD	38	44	
Precision		61.22%	66.67%	

Accuracy= 63.42%, Kappa = 0.268

Figure 5: Fusion of Top Three Results

MV using RF and 59.76% for ABR using RF) and fused their label using the late fusion method. The fusion provides the best accuracy of 63.42%. The confusion matrix along with precision, recall and Kappa is shown in Figure 5.

Table 4: Classification Results using Affective Behaviour

Method	Blind	LDA	DT	INN	SVM	RF
Segment Level	50.12	48.68	40.78	51.67	50.64	52.70
Majority Vote	50.00	50.61	40.24	57.93	54.27	<b>58.54</b>
ABR	50.00	54.27	50.61	54.27	54.27	<b>59.76</b>

## 7. Discussion

Statistical analysis showed that non-AD subjects' speech segments have a higher mean value (0.1569) of classification score for Anger than AD subjects (0.1285), the difference is statistically significant ( $p < 0.01$ ). This suggests that the non-AD subjects expressed characteristics of Anger in their speech more than the AD subjects. Non-AD speech segments also have a higher mean value (0.1174) of classification score for Disgust than AD subjects (0.1026), similarly suggesting that non-AD subjects expressed this emotion in their speech more than AD subjects. This is likely due to the fact that the AD subjects usually have lower voice volume, speech rate and pitch than non-AD subjects, while Anger and Disgust emotions are associated with high voice volume, speech rate and pitch.

The non-AD subjects' speech segments have a significantly lower mean value (0.0815) of classification score for Sadness than AD subject (0.1297), suggesting that AD participants expressed speech with characteristics of Sadness more than non-AD subjects. For expression of Boredom, Fear and Happy, the differences in classifier scores are not statistically significant, which suggests that either AD and non-AD subjects can express those emotions equally in their speech, or that the model is not discriminating enough to detect those expressions. It is also noted that the emotion recognition system is more certain about the emotions of AD (mean entropy of 2.4413) than non-AD (mean entropy of 2.4750) participant, and this difference though quite small is statistically significant at the  $p < 0.05$  level. As regards with the AD recognition results, using scores of the emotion recognition system and ABR as input features for the AD classifier we were able to detect AD speech with an accuracy of 63.42%. While this scores is below state-of-the-art AD classifiers (Haider et al., 2020), we note that the emotion recognition system evaluated on *emoDB* dataset

is also around 69.72% accurate. A speech dataset annotated for the emotions of elderly people and people with AD could conceivably improve the quality of the input features, and the performance of our emotion-based approach to AD recognition. It is also noted that the emotion recognition model was trained on a dataset (*emoDB*) recorded in a different language (German) to the one of the Pitt data (English). While the annotation quality of *emoDB* is higher than other datasets such as (Haq and Jackson, 2009; Costantini et al., 2014), it is possible that an affect recognition system trained directly on English data might improve the results.

## 8. Conclusion

In conclusion, we found that there are differences in (automatically) inferred affective behaviours regarding expressions of Sadness, Anger and Disgust among AD and non-AD subjects. Although these results need further study, they suggest, in agreement with the incipient literature on this topic, that AD speakers exhibit a deficit in the expression of those emotions reflected on voice volume, speech rate and pitch. The proposed Affective Behaviour Representation (ABR) and emotion classification scores are able to predict the AD with an accuracy of 63.42%. A limitation of this study which should be addressed in future work is the mismatch between the dataset used to generate the features for AD recognition (*emoDB*) and the *Pitt Corpus* (in which these features were used). This includes the facts that (unlike *emoDB*) the *Pitt Corpus* was not explicitly designed to elicit emotions, that the two datasets were recorded under different acoustic conditions and demographics, and that they are in different languages. In future work, we intend to manually annotate the Pitt corpus for emotions, and train an affect recognition system based on this augmented dataset to assess the effect of this model on AD recognition accuracy. The affect recognition system along with ABR script is made available to the research community through a git repository <sup>3</sup>.

## 9. Acknowledgements

SdIFG and PA are supported by the Medical Research Council (MRC). This research is also funded by the European Union's Horizon 2020 research programme, under grant agreement No. 769661, towards the SAAM project. We acknowledge B. MacWhinney (Dementia-Bank) for hosting and sharing the *Pitt Corpus* database.

## 10. Bibliographical References

- American Psychiatric Association. (2000). Delirium, dementia, and amnesic and other cognitive disorders. In American Psychiatric Association, editor, *Diagnostic and Statistical Manual of Mental Disorders, Text Revision (DSM-IV-TR)*, chapter 2. Arlington, VA, fourth edition.
- Becker, J. T., Boiler, F., Lopez, O. L., Saxton, J., and McGonigle, K. L. (1994). The Natural History of Alzheimer's Disease. *Archives of Neurology*, 51(6):585.

<sup>3</sup><https://git.ecdf.ed.ac.uk/fhaider/emotion2dementia.git>

- Bondi, M. W., Salmon, D. P., and Kaszniak, A. W. (1996). The neuropsychology of dementia. In *Neuropsychological assessment of neuropsychiatric disorders.*, pages 164–199. Oxford University Press, New York, NY, US, 2 edition.
- Bucks, R. S. and Radford, S. A. (2004). Emotion processing in alzheimer’s disease. *Aging & mental health*, 8(3):222–232.
- Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., and Weiss, B. (2005). A database of german emotional speech. In *Proceedings of the ninth European Conference on Speech Communication and Technology*, pages 1516–1520.
- Corey Bloom, J. and Fleisher, A. (2000). The natural history of alzheimer’s disease. *Dementia*, 34:405–15.
- Costantini, G., Iaderola, I., Paoloni, A., and Todisco, M. (2014). Emovo corpus: an italian emotional speech database. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC), LREC 2014*, pages 3501–3504. European Language Resources Association (ELRA).
- Dukart, J., Schroeter, M. L., Mueller, K., Initiative, A. D. N., et al. (2011). Age correction in dementia-matching to a healthy brain. *PLoS one*, 6(7):e22193.
- Eyben, F., Weninger, F., Groß, F., and Schuller, B. (2013). Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 835–838. ACM, Association for Computing Machinery.
- Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., et al. (2016). The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202.
- Haider, F., de la Fuente, S., and Luz, S. (2020). An assessment of paralinguistic acoustic features for detection of alzheimer’s dementia in spontaneous speech. *IEEE Journal of Selected Topics in Signal Processing*, 14(2):272–281.
- Han, K.-H., Zaytseva, Y., Bao, Y., Pöppel, E., Chung, S. Y., Kim, J. W., and Kim, H. T. (2014). Impairment of vocal expression of negative emotions in patients with alzheimer’s disease. *Frontiers in aging neuroscience*, 6:101.
- Haq, S. and Jackson, P. (2009). Speaker-dependent audio-visual emotion recognition. In *Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP)*, pages 53–58, Sept.
- Hart, R. P., Kwentus, J. A., Taylor, J. R., and Harkins, S. W. (1987). Rate of forgetting in dementia and depression. *Journal of Consulting and Clinical Psychology*, 55(1):101–105.
- Horley, K., Reid, A., and Burnham, D. (2010). Emotional prosody perception and production in dementia of the alzheimer’s type. *Journal of Speech, Language, and Hearing Research*, 53:1132–1146.
- Johnstone, T. and Scherer, K. R. (2000). Vocal communication of emotion. *Handbook of emotions*, 2:220–235.
- Juslin, P. N. and Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological bulletin*, 129(5):770.
- Kirshner, H. S. (2012). Primary Progressive Aphasia and Alzheimer’s Disease: Brief History, Recent Evidence. *Current Neurology and Neuroscience Reports*, 12(6):709–714.
- Landes, A. M., Sperry, S. D., and Strauss, M. E. (2005). Prevalence of apathy, dysphoria, and depression in relation to dementia severity in alzheimer’s disease. *The Journal of neuropsychiatry and clinical neurosciences*, 17(3):342–349.
- Leyhe, T., Reynolds III, C. F., Melcher, T., Linnemann, C., Klöppel, S., Blennow, K., Zetterberg, H., Dubois, B., Lista, S., and Hampel, H. (2017). A common challenge in older adults: Classification, overlap, and therapy of depression and dementia. *Alzheimer’s & dementia*, 13(1):59–71.
- Lopes, P. N., Brackett, M. A., Nezlek, J. B., Schütz, A., Sellin, I., and Salovey, P. (2004). Emotional intelligence and social interaction. *Personality and social psychology bulletin*, 30(8):1018–1034.
- Lopez-de Ipiña, K., Faundez-Zanuy, M., Solé-Casals, J., Zelarín, F., and Calvo, P. (2016). Multi-class Versus One-Class Classifier in Spontaneous Speech Analysis Oriented to Alzheimer Disease Diagnosis. In Esposito et Al., editor, *Recent Advances in Nonlinear Speech Processing*, volume 48, pages 63–72. Springer International Publishing.
- Roberts, V. J., Ingram, S. M., Lamar, M., and Green, R. C. (1996). Prosody impairment and associated affective and behavioral disturbances in alzheimer’s disease. *Neurology*, 47(6):1482–1488.
- Seidl, U., Lueken, U., Thomann, P. A., Kruse, A., and Schröder, J. (2012). Facial expression in alzheimer’s disease: impact of cognitive deficits and neuropsychiatric symptoms. *American Journal of Alzheimer’s Disease & Other Dementias®*, 27(2):100–106.
- Starkstein, S. E., Ingram, L., Garau, M., and Mizrahi, R. (2005). On the overlap between apathy and depression in dementia. *Journal of Neurology, Neurosurgery & Psychiatry*, 76(8):1070–1074.
- Taler, V., Baum, S. R., Chertkow, H., and Saumier, D. (2008). Comprehension of grammatical and emotional prosody is impaired in alzheimer’s disease. *Neuropsychology*, 22(2):188.
- Testa, J., Beatty, W., Gleason, A., Orbelo, D., and Ross, E. (2001). Impaired affective prosody in ad: Relationship to aphasic deficits and emotional behaviors. *Neurology*, 57(8):1474–1481.



# Individual Mandibular Motor Actions Estimated from Speech Articulation Features

Andrés Gómez-Rodellar<sup>1</sup>, Athanasios Tsanas<sup>1</sup>, Pedro Gómez-Vilda<sup>2</sup>, Agustín Álvarez-Marquina<sup>2</sup>, Daniel Palacios-Alonso<sup>3</sup>

<sup>1</sup>Usher Institute, Edinburgh Medical School, University of Edinburgh, UK

<sup>2</sup>Neuromorphic Speech Laboratory, Center for Biomedical Technology, Universidad Politécnica de Madrid, Campus de Montegancedo, s/n, 28223 Madrid, Spain

<sup>3</sup>Escuela Técnica Superior de Ingeniería Informática - Universidad Rey Juan Carlos, Campus de Móstoles, Tulipán, s/n, 28933 Móstoles, Madrid, Spain

andres.gomez@ctb.upm.es, Athanasios.Tsanas@ed.ac.uk, pedro@fi.upm.es, agustin@junipera.datsi.fi.upm.es, daniel.palacios@urjc.es

## Abstract

This study aims to compare the articulation characteristics of a person with Parkinson's Disease (PD) with the articulation characteristics of a healthy person on the neuromotor principles of speech production. The study methodology is based on the recording of speech as a vehicular signal accompanied by other multimodal traits associated, as the surface Electromyography in the masseter and the acceleration in the chin. Diadochokinetic exercises are used to elicit specific recordings with special interest for the study. The two first formants derived from speech acoustic analysis are transformed into a kinematic velocity which is compared with the dynamic velocity obtained from accelerometry. Comparisons in terms of Jensen-Shannon Divergence (JSD) between amplitude distributions of formant-derived Kinematic Correlates and accelerometry-derived Dynamic Correlates from speech of a Control Subject (CS) and a PD patient allow to appreciate a different behavior in both types of correlates. In the case of the CS the similarity between formant-derived and acceleration derived correlates is large, but small in the PD case. The similarity between the CS and the PD correlates in terms of the formant-derived kinematics is medium, and very small in terms of acceleration-derived dynamics. These comparisons may help in establishing a new relationship between speech acoustic analysis and neuromuscular dynamics in speech production.

**Keywords:** Speech Production, Neuromotor Dynamics, Neurodegenerative Disorders, Sensors, Multimodal Signal Analysis

## 1. Introduction

The study of speech in neurodegenerative diseases, such as Parkinson's Disease (PD) has been conceived under two different but complementary approaches.

The first approach focuses on extracting features derived from the statistical properties of first level observables as positions and speeds combined with powerful machine learning methods. These are able of producing significant classification results regarding the presence and prediction of severity degree of dysfunctions in movement involving axial correlates in PD as in handwriting and speech production, of high value in disease monitoring and tracking (Carmona et al., 2018).

The second approach in the field of speech production is based on mechanistic models in the study of the fine structure of small movements in the main muscles involved in speech production, namely the masseter, the tongue and the larynx complex. This approach has the potential of providing new insights into the physiology of neuropathological speech production (Yunusova, Weismer and Lindstrom, 2011), (Gómez et al., 2019a).

The study is divided in five main sections. Section 2 is devoted to define the objectives of the study. Section 3 describes the characteristics of the subjects participating in the study, the signals recorded and the estimation methods used to obtain acoustic and dynamic correlate, and the Mutual Information distance proposed to establish comparisons. The results of the intra- and inter-subject are presented and discussed in Section 4. The study's key findings are summarized in Section 5.

## 2. Objectives

The aim of the present study is to investigate the neuromotor activity on the masseter as one of the main muscles active in speech articulation. The masseter is a major muscle which experiences a strong neuro-electrical activity. It is composed of individual motor fiber discharges, which can be easily detected at the skin level as surface Electro-Myographic Signals (sEMG). The resulting muscle contractions may be estimated using 3D Accelerometry (3DAcc), therefore there is a physiological reference to validate other measurements quantifying acoustic characteristics, which are indirectly inferred from the speech signal.

The main acoustic correlates found in speech regarding articulation are formants, which may be seen as frequencies enhanced by the resonant characteristics of the vocal tract and well established acoustic features in describing speech articulation dynamics (Huang, et al., 2001). Formants may be estimated from the inverse filtering of the speech signal (Alku et al., 2019). Speech formants reflect the modification of the resonant cavities of the vocal tract, especially in the oral cavity, where the masseter muscle is one main actor involved in producing open and close vowels. The first step in using speech as a remote monitoring tool is to find dynamic correlates on the speech signal, which can be validated using 3D acceleration. In a further step the objective is to find biomarkers to assess how PD patients differ from a healthy control population helping in early diagnosis.

The focus of this study is to confirm that signals produced exclusively from speech can be used as correlates to infer neuromotor activity in the speech production complex.

This would facilitate greatly the analysis of neurodegenerative disorders from speech correlates at a neurophysiological level using only speech as the vehicular measurement (Orozco-Arroyabe et al., 2016).

### 3. Materials and Methods

For this purpose, a joint jaw-tongue biomechanical model as the one illustrated in Figure 1 will be used. The jaw-tongue biomechanical system is modeled as a third-order lever system with a lumped mass load concentrated in the reference point  $P_{JT}$   $\{x_r, y_r\}$  (Hannam et al., 2008). Harmonic oscillation  $\{\Delta x_r, \Delta y_r\}$  around the fulcrum (F: attachment to the skull) is assumed under forces acting on this system. A very descriptive kinematic correlate of the jaw-tongue neuromotor activity is the Absolute Kinematic Velocity (AKV)  $|v_r|$  of the reference point  $P_{JT}$ :

$$|v_r| = \left[ \left( \frac{dx_r}{dt} \right)^2 + \left( \frac{dy_r}{dt} \right)^2 \right]^{\frac{1}{2}} \quad (1)$$

The statistical distribution of the AKV will contribute valuable information in characterizing unstable articulation, as explained in the sequel.

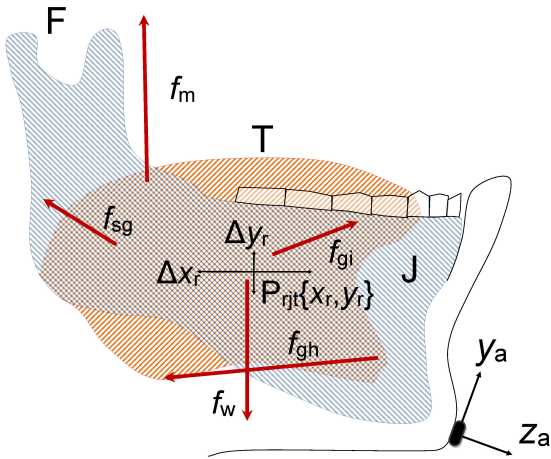


Figure 1. Jaw-Tongue Model. F: Fulcrum; T: Tongue; J: Jaw bone; H: Hyoid bone;  $f_{sg}$ : stylo-glossus force;  $f_m$ : masseter force;  $f_{gi}$ : glosso-intrinsic forces;  $f_{gh}$ : genio-hyoid force;  $f_w$ : gravity;  $y_a$ ,  $z_a$ : accelerometer tangential and normal components;  $\Delta x_r$ ,  $\Delta y_r$ : horizontal and vertical displacements of the reference point ( $P_{JT}$ ) in the sagittal plane.

The signals used in this study are the speech, the sEMG taken at the masseter and the three-channel 3DAcc taken at the chin of a normative 30 year-old male subject (CS) and a PD patient (male, 77 year-old, in stage 2 of the H&Y scale). The sEMG activity is captured by a pair of surface electrodes at the condilo-maxillary attachments of the masseter, the 3DAcc is taken by an accelerometer placed at the chin end, and the speech signal is captured by a cardioid high-quality clip microphone on the chest at about 20 cm away from the radiation point (mouth). The experimental setup for the data collection is illustrated in Figure 2.

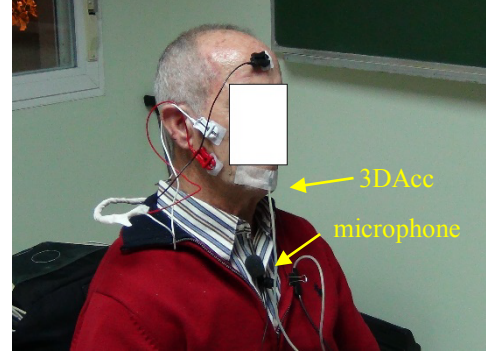


Figure 2. Instrumentation settings for the simultaneous recording of speech (microphone), sEMG (red and white: active electrodes; black: reference) and 3DAcc.

The current experimental fixture is based in a signal acquisition architecture on a BioPac MP 150 platform with the corresponding modules: one analog channel for speech acquisition at 50 kHz and 16 bits, one analog channel for sEMG at 1,562 Hz and 16 bits, and three analog channels from a  $\pm 3.5g$  accelerometer for 3DAcc sampled at 1,562 Hz and 16 bits. The speech signal is downsampled at 16 kHz, and the sEMG and 3DAcc are downsampled at 500 Hz. The sEMG is processed by a 50 Hz stop-band filter to cancer power line spurious noise. Signal acquisition is synchronous for all channels involved.

An example of the set of measurements resulting from the diadochokinetic exercise consisting in the continuous non-stop repetition of the diphthong [ay] produced by a normative male subject is shown in Figure 3.

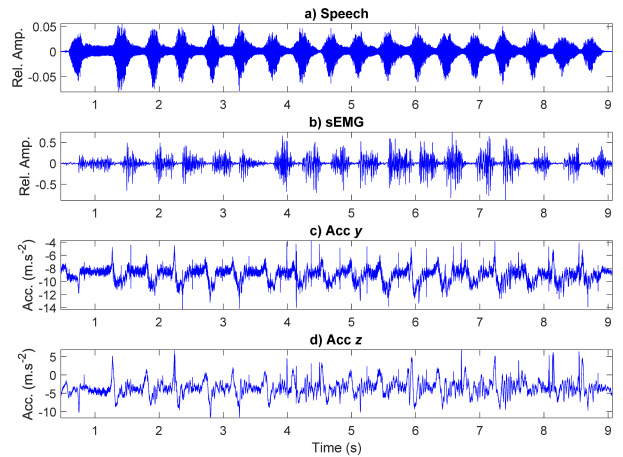


Figure 3. Multimodal signal measurements from the repetition of diphthong [ay] from a CS: a) speech recording; b) sEMG signal on the masseter; c) tangential acceleration component; d) normal acceleration component.

The recorded signals are processed accordingly to their respective nature and purpose, as follows:

- The 3DAcc is referenced to the center of masses of the joint jaw-tongue complex (Gómez et al., 2019b) and integrated to produce the Absolute Kinematic Velocity (AKV) of this point as the square root of each component squares.

- The speech signal is inverted to estimate the first two formants (Alku et al, 2019). The contributions of each formant are differentiated to produce the Absolute Formant Velocity (AFV) as the square root of each component squares (Gómez et al., 2019c).
- Although sEMG has been recorded simultaneously it will not be considered in the present study.

The AKV and AFV signals and distributions for the CS and PD patient are given in Figure 3 as an example. The AKV and especially the AFV represent the actions involved in the muscular activity of the masseter, showing paired lumped maxima which are related with agonist/antagonist activations. Besides, as it may be seen at first sight both AKV and AFV present some resemblance, which may be more evident comparing their normalized distributions.

The AKV and especially the AFV represent the actions involved in the muscular activity of the masseter, showing paired lumped maxima which are related with agonist/antagonist activations. Besides, as it may be seen at first sight both AKV and AFV present some resemblance, which may be more evident comparing their normalized distributions. We quantified the difference between distributions using the Jensen-Shannon Divergence (JSD) (Cover and Thomas, 2006) between the intra-speaker AKV and AFV from their normalized histograms (pdfAKV and pdfAFV):

$$D_{JS} = \frac{D_{KL}(p(x)|m(x)) + D_{KL}(q(x)|m(x))}{2} \quad (2)$$

where  $m(x)$  is the average of  $p(x)$  (pdfAKV) and  $q(x)$  (pdfAFV), and  $D_{KL}$  is a modified divergence estimate based on Kullback-Leibler's Divergence (Salicrú et al, 1994, Georgiou and Lindquist, 2003):

$$D_{KL}(p|q) = \int_0^{\infty} p(x) \log \left\{ \frac{p(x)}{q(x)} \right\} dx \quad (3)$$

The JSD is symmetrical with respect to  $p(x)$  and  $q(x)$ , and it may be considered a normalized and limited distance to the interval  $[0, 1]$ .

#### 4. Results and Discussion

In the present study a comparison between the articulation capabilities of a CS and a PD patient is to be established in terms of the mutual JSD on their respective pdfAKV and pdfAFV. The results of the comparison are given in Table 1.

Comparisons	JSD
1. pdfAKVCS vs pdfAFVCS	0.1690
2. pdfAKVPD vs pdfAFVPD	0.5545
3. pdfAKVCS vs pdfAKVPD	0.3685
4. pdfAFVCS vs pdfAFVPD	0.6501

Table 1 . Jensen-Shannon Divergence between self and cross-speaker distributions in comparing a male CS with a male PD patient.

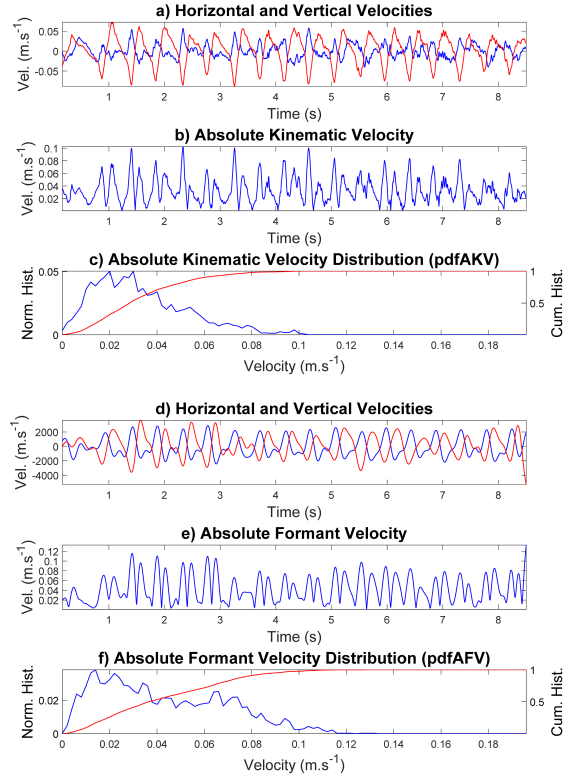


Figure 4. a) Horizontal (blue) and vertical (red) velocities of the jaw-tongue reference point from 3D accelerometry in the chin, after rotation and integration (sagittal plane); b) AKV from the horizontal and vertical components; c) Normalized (blue) and cumulative (red) distributions of the AKV. d) First difference of the first (blue) and second (red) formants estimated from speech inversion; e) AFV estimated from the first differences of both formants applying kinematic compensation (Gómez et al., 2019b); f) Normalized (blue) and cumulative (red) distributions of the AFV. These signals have been obtained on the utterance from a CS shown in Figure 1.

These comparisons allow an interesting interpretation taking into account that when  $0 < JSD < 0.25$  a distance is considered small, expressing a certain similarity between the distributions under comparison; when  $0.25 < JSD < 0.5$  the distance will be considered medium, and a small dissimilarity would be observed, and if  $JSD > 0.5$  the distance will be considered large and the distributions are considered dissimilar. The discrimination thresholds have been determined heuristically.

It is observed that the distance between pure kinematic and formant-based distributions is relatively similar for the CS (comparison 1), but it is largely dissimilar for the PD patient (comparison 2). Apparently, this considerable deviation could be due to a very much reduced extension of the formant span in the case of the PD patient compared with the patient's kinematic distribution. This is also observed in comparison 3, where the JSD between the kinematic distributions between CS and PD pdfAKV's is larger than in the CS, but not as large as in the JSD between both speakers pdfAFV's (comparison 4).

In summary, it seems that the reduction in the kinematic activity of the PD patient is not as strong as the one

observed in his speech derived from the formant dynamic distribution. Possibly some other factors influencing resonances in the vocal tract may play a secondary role in the reduction of the formant dynamic span.

These results were obtained on data from a case study involving a CS and a PD patient, and are limited in their reach and scope, therefore a deeper investigation building in these pilot findings is needed to verify these preliminary results on a fair-size balanced dataset including male and female CS and PD patients.

## 5. Conclusions

The present work has used a pdfAKV derived from the jaw-tongue movement estimated when producing speech performing certain diadochokinetic exercises. The amplitude distribution of this correlate may be used in characterizing the kinematic activity of the masseter. Some tentative insights may be concluded from these exploratory results:

- An AFV may be derived from formant dynamics, which is strongly related with the AKV, connecting relating biomechanical and acoustic correlates.
- The similarity between intra-speaker and inter-speaker kinematic activity may be estimated using Information Theory methods.
- Important differences may be observed between the CS and the PD subjects involved in this experimentation.
- These results need to be confirmed by more estimations from a populated database in a future study.

## 6. Acknowledgments

This study has been supported by grants TEC2016-77791-C4-4-R (MINECO, Spain) and CENIE\_TECA-PARK\_55\_02 INTERREG V-A Spain – Portugal (POCTEP).

## 7. References

- Alku, P., Murtola, T., Malinen, J., Kuortti, J., Story, B., Airaksinen, M., Salmi, M., Vilkmán, E., Geneid, A. (2019). OPENGLot-An open environment for the evaluation of glottal inverse filtering. *Speech Communication*, 107: 38-47.
- Carmona, C., Ferrer, M. A., Gómez, P., Van Geemert, A. W. A., Plamondon, R. (2018). A common framework to evaluate Parkinson's disease in voice and handwriting. *Proc. ICPRAI*: 795-799.
- Cover, T. M and Thomas, J. A. (2006). *Elements of Information Theory*. Wiley Interscience.
- Endres, D. M. and Schindelin, J. E., (2003). A New Metric for Probability Distributions. *IEEE Trans. on Information Theory*, 49(7): 1858-1860.
- Georgiou T. and Lindquist, A., (2003). Kullback-Leibler Approximation of Spectral Density Functions. *IEEE Trans. on Information Theory*, 49(11): 2910-2917.
- Gómez, A., Tsanas, A., Gómez, P., Palacios, D., Álvarez, A., Martínez, R. (2019a). A Neuromechanical Model of Jaw-Tongue Articulation in Parkinson's Disease Speech. *Proc. of MAVEBA 2019*, Firenze University Press: 25-28.
- Gómez, P., Gómez, A., Ferrandez, J. M., Mekyska, J., Palacios, D., Rodellar, V., Álvarez, A., Smekal, Z., Eliasova, I., Kostalova, M., Rektorova, I. (2019b). Neuromechanical Modelling of Articulatory Movements from Surface Electromyography and Speech Formants. *International Journal of Neural Systems*, 29(2): 1850039.
- Gómez, P., Mekyska, J., Gómez, A., Palacios, D., Rodellar, V., Álvarez, A. (2019c). Characterization of Parkinson's disease dysarthria in terms of speech articulation kinematics. *Biomedical Signal Processing and Control*, 52: 312-320.
- Hannam, A. G., Stavness, I., Lloyd, J. E., Fels, S. (2008). A dynamic model of jaw and hyoid biomechanics during chewing. *J. Biomechanics*, 41: 1069-1076.
- Huang, X., Acero, A., Hsiao-Wuen, H. (2001). *Spoken Language Processing*. Prentice-Hall.
- Orozco-Arroyabe, J. R., Hönl, F., Arias-Londoño, J. D. Vargas-Bonilla, F., Daqrouq, K., Skodda, S., Ruz, J., Nöth, E., (2016). Automatic detection of Parkinson's Disease in running speech spoken in three different languages. *Journal of the Acoustical Society of America*, 139: 481-500.
- Salicrú, M., Morales, D., Menéndez, M. L., Pardo, J. (1994). On the Applications of Divergence Type Measures in Testing Statistical Hypotheses, *Journal of Multivariate Analysis*, 51(2): 372-391.
- Yunusova, Y., Weismer, G. and Lindstrom, M. J. (2011). Classifications of Vocalic Segments From Articulatory Kinematics: Healthy Controls and Speakers with Dysarthria, *Journal of Speech, Language and Hearing Research*, 54: 1302-1311.

# Digital Eavesdropper – Acoustic Speech Characteristics as Markers of Exacerbations in COPD Patients

Julia Merkus<sup>1</sup>, Ferdy Hubers<sup>1,2</sup>, Catia Cucchiari<sup>1</sup>, Helmer Strik<sup>1,2,3</sup>

<sup>1</sup> Centre for Language and Speech Technology (CLST), <sup>2</sup> Centre for Language Studies (CLS),

<sup>3</sup> Donders Institute for Brain, Cognition and Behaviour

Radboud University Nijmegen

{J.Merkus, F.Hubers, C.Cucchiari, W.Strik }@let.ru.nl

## Abstract

Speech production is a complex human activity that may be affected by both physical and cognitive disorders. Since deviations in speech production can potentially reveal diseases apparently unrelated to speech, speech-based early detection is currently a topical issue. In the present research, we study whether changes in speech can indicate an upcoming exacerbation in patients with chronic obstructive pulmonary disease (COPD). Identifying potential early markers of exacerbation could help develop valuable patient remote monitoring systems. In our study, we focus on acoustic parameters and study whether they are suitable for automatic detection of exacerbations in COPD patients by investigating a) which aspects of speech differ between COPD patients and healthy speakers, and b) which aspects differ between COPD patients in exacerbation and stable conditions. Read speech of 9 stable COPD patients and 5 healthy controls (I), and 9 COPD patients in exacerbation vs. stable conditions (II) was analysed and compared. Results showed that harmonics-to-noise ratio (HNR), shimmer, vowel duration and the number of (non-linguistic) inhalations per syllable are potential markers that could be employed in remote monitoring systems. Further research is needed to examine the validity of the results for other types of speech and larger sample sizes.

**Keywords:** chronic obstructive pulmonary disease (COPD), exacerbations, early markers, acoustic speech characteristics, remote monitoring, e-health

## 1. Introduction

Speech production is a complex and demanding human activity involving cognition and a carefully coordinated effort of all bodily parts involved in this process. A change (e.g. deterioration) in one or multiple processes involved in speech production (e.g. cognition or motor skills) might thus affect the resulting speech signal. As a result, various diseases, even those that at first sight might seem unrelated to speech, like multiple sclerosis, Parkinson's and chronic obstructive pulmonary disease (COPD), can manifest themselves in speech production. This awareness has spawned a whole body of research aimed at early detection and diagnosis based on speech analysis. This became very clear at the recent annual conference of the international speech communication association (ISCA), Interspeech 2019, that was held in Graz (Austria) in September 2019 ([https://www.isca-speech.org/archive/Interspeech\\_2019/](https://www.isca-speech.org/archive/Interspeech_2019/)). Special sessions such as 'Medical Applications and Visual ASR', 'Speech and Voice Disorders 2', 'Speech and Audio Characterization and Segmentation', and 'Speech and Language Analytics for Medical Applications' envisaged many presentations on speech-based early detection or diagnosis of different 'diseases', such as dementia, Alzheimer's (Chen, Zu and Jieping, 2019), dysarthria (Shor et al., 2019), amyotrophic lateral sclerosis (Gutz et al., 2019) and depression level (Rutowski et al., 2019).

An important advantage is that speech analysis is generally non-invasive, efficient, and easy to apply. In addition, the current ubiquity of voice-based services would seem to make the collection of speech recordings increasingly easier, provided the right measures are taken to ensure privacy and security. In the current research, we study whether changes in the properties of the speech signal might also be useful for early detection of an exacerbation (lung attack) for COPD patients.

COPD is an umbrella term used to describe progressive lung diseases characterised by airflow limitation. It is diagnosed by measuring the extent of the airflow limitation. In most cases, spirometry is used to measure the lung function (GOLD, 2019). Spirometry is more reliable than clinical descriptions, since COPD is a heterogeneous disease in its clinical expression (Postma, Bush and Van den Berge, 2015). Stable COPD is interrupted by episodes or exacerbations during which the respiratory symptoms acutely worsen (GOLD, 2019). Exacerbations are often recognised at a late stage, which delays the treatment (Hurst et al., 2010). It is of great importance to diagnose the patient quickly and correctly, because early treatment shortens the duration and seriousness of the exacerbation and could prevent COPD patients from being hospitalised (Trappenburg et al., 2011). COPD patients would benefit from a smart, automated, remote system which is able to correctly identify changes in speech acoustics that indicate an upcoming exacerbation. However, previous studies on the use of e-monitoring to help manage patients with COPD have not yet led to consensus regarding the most suitable acoustic parameters to detect exacerbations (Bolton et al., 2010).

The current research aims to identify those acoustic speech characteristics that mark an upcoming exacerbation in COPD patients. For this purpose, an exploratory analysis was conducted of the differences in acoustic measures for stable COPD patients and healthy controls.

In the remainder of this paper we first discuss relevant background research on COPD and e-health and introduce the current study (Section 2). In Section 3 we present the methods adopted in this study. Subsequently, we report the results in Section 4. Finally, in Section 5, we discuss our findings, relate them to those of previous studies, and suggest possible avenues for future research.

## 2. Research background

### 2.1 Chronic Obstructive Pulmonary Disease (COPD)

According to the guidelines provided by the Global Initiative for Chronic Obstructive Lung Disease (GOLD, 2019), the official definition of COPD is “a common, preventable and treatable disease that is characterised by persistent respiratory symptoms and airflow limitation that is due to airway and/or alveolar abnormalities usually caused by significant exposure to noxious particles or gases” (p. 2).

The prevalence of COPD worldwide is estimated at roughly 12%, but the percentage differs greatly between different subgroups (López et al., 2014). Taking into account the three million annual deaths globally, COPD is currently the fourth leading cause of death in high-income countries and it is expected to be the third leading cause in 2020 due to a higher life expectancy and increasing air pollution (Buist et al., 2007; GOLD, 2019; Postma et al., 2015). However, the lung disease has been overlooked and neglected for a long time by both the public and the pharmaceutical industry. This neglect might have been caused in part by the assumption that COPD is a self-inflicted health condition caused by smoking. Although smoking is the leading cause of COPD in high-income countries, over 15% of the patients are nonsmokers (Buist et al., 2007). The four greatest predictors of COPD are years and intensity of smoking, age, sex and body mass index (Lopez et al., 2014). COPD is associated with an economic burden, since the disease accounts for approximately 55% of the costs for respiratory diseases in Europe (GOLD, 2019).

Individuals suffering from COPD generally show a variety of symptoms, commonly including shortness of breath, tightness on the chest and coughing (with mucus). Stable COPD is interrupted by episodes or exacerbations during which the respiratory symptoms acutely worsen (GOLD, 2019). During exacerbations, the peripheral airway limitation causes gas to get trapped during expiration. This leads to hyperinflation, which is associated with a limited inspiratory capacity and increased dyspnoea. The morbidity increases with age, because other concomitant chronic conditions, such as diabetes and cardiovascular diseases, often interfere with COPD management and might compromise the patient’s health even more. This causes COPD patients to be hospitalised frequently (GOLD, 2019). Currently, COPD cannot be reversed or cured. However, different types of medication provide patients with symptomatic relief and an improved quality of life. During exacerbations, patients are often treated with either Prednisone or oxygen treatment (Soriano, Zielinski and Price, 2009).

COPD patients generally show an abnormal swallowing pattern where they swallow more often and more distinctively (Robinson et al., 2011). Moreover, a study by Vyshedskiy and Murphy (2016) on the acoustic biomarkers of COPD showed that there are measurable differences between the lung sound patterns of COPD patients compared to age matched controls. The breathing difficulties could result in a deviant intensity, a lack of intonation, decreased phonation time for sustained vowels,

and a deviant pitch, depending on the compensatory strategy (Constantinescu et al., 2010).

### 2.2 E-health

There are three common types of e-practice according to the American Speech-Language-Hearing Association (ASHA, 2014): synchronous, client-interactive services (conducted with video and audio connections in real time), asynchronous, store-and-forward services (data are obtained and transmitted for professional interpretation) and hybrid services (combination of the first two). However, a fourth type of e-health services, ‘remote patient monitoring’, is more suitable for remotely diagnosing or monitoring patients instead of treating them. This includes a sensor or tracking device, which analyses several parameters related to the patient’s condition (Ong et al., 2016). Any changes in these parameters could result in an automatic warning for the patient and/or medical professional. Two potential benefits of remote patient monitoring are the restoration of a patient’s sense of autonomy to a certain extent, and the decrease in medical costs (Klersy et al., 2014).

Previous studies regarding language and speech pathologies have provided a broad range of outcome measures, which are suitable for the detection or monitoring of pathological speech.

### 2.3 The Present Study

Exacerbations are often recognised at a late stage, which delays the treatment (Hurst et al., 2010). It is of great importance to diagnose the patient quickly and correctly, because early treatment shortens the duration and the seriousness of the exacerbation and could prevent COPD patients from being hospitalised (Trappenburg et al., 2011). COPD patients would benefit from a smart, automated remote system which is able to correctly identify changes in speech acoustics that indicate an upcoming exacerbation. However, previous studies on the use of e-monitoring to help manage patients with COPD have not yet led to consensus regarding the most suitable acoustic parameters to detect exacerbations (Bolton et al., 2010).

Medical professionals from the Department of Pulmonary Diseases of Dekkerswald (Radboud University Medical Centre) have indicated that they are able to estimate the condition of a patient who suffers from COPD by listening to their speech. This means they believe the speech of COPD patients during stable periods differs from their speech during exacerbations, mainly with respect to their breathing pattern and vocal quality. The information regarding the acoustic speech characteristics of COPD patients could be of great value if the professionals’ observations are correct. The acoustic markers could be implemented in speech pathology recognition software to prematurely identify signs of deterioration and warn the patient (and health care professional).

The current research provides an exploratory analysis of the differences in acoustic measures for stable COPD patients and healthy controls. The following research questions have been formulated for this study:

1. Which (acoustic) measures extracted from read speech differ for COPD speakers in stable condition and healthy speakers?

2. In which aspects does the speech of COPD patients during an exacerbation differ from the speech of COPD patients during stable periods?

2.1 Which (acoustic) measures extracted from read speech differ for COPD speakers during an exacerbation and COPD speakers in stable condition?

2.2 Which (acoustic) measures extracted from sustained vowels differ for COPD speakers during an exacerbation and COPD speakers in stable condition?

### 3. Methods

#### 3.1 Participants

The speech of eleven native speakers of Dutch (seven males, four females) was recorded twice in a treatment room of the lung department of the Radboud University Medical Centre in the period from August 2016 until April 2017. All participants had officially been diagnosed with COPD. The participants were hospitalised due to an exacerbation and they had to stay in the hospital for two to 23 days ( $M = 8.82$ ,  $SD = 6.11$ ). The participants were requested to participate in the experiment after first receiving urgent care for their exacerbation. Patients suffering from additional lung diseases were excluded from participation. Furthermore, five recordings of five healthy, adult speakers (four males, one female) from the Spoken Dutch Corpus ('Corpus Gesproken Nederlands', CGN) (Oostdijk, 2000) were selected in order to compare (characteristics of) their speech with the speech of the COPD patients. These five recordings were obtained between 1998 and 2004. Detailed participant information, such as the COPD stage, number of previous exacerbations, age, gender and years since diagnosis, is missing due to the retrospective nature of the research.

#### 3.2 Materials

The recordings were made using two Relitech microphones and each recording consisted of two parts: a sustained vowel and read speech. Sustained vowels have proven to be very effective to efficiently distinguish between healthy and pathological speakers due to the clean signal, which allows for a better extraction of voice features (Boyanov and Hadjitodorov, 1997). However, sustained vowels do not commonly occur in natural speech. In an attempt to approach more natural situations while at the same time maintaining some control over the experiment, we decided to use read speech (Maryn and Roy, 2012).

1. The first part represented a sustained phonation of the vowel [a:]. The participants were free to choose a comfortable vocal intensity.

2. The second part consisted of a reading of the phonetically balanced story 'De Koning' (515 words) (Bomans, 1946). This story contains existing, highly frequent words and the sentences differ in length (Haasnoot, 2012). The data collector selected part of the story to read (part 1: 90 – 140 words; part 2: 111 - 259 words). The parts differed slightly between subjects, because a few patients were physically unable to finish the whole part. The patients read different parts during their first and second recording to avoid a learning effect.

The database for the pathological speakers thus contained eleven recordings of a sustained vowel during an exacerbation (i), eleven recordings of a sustained vowel during stable COPD (ii), eleven recordings in which the patient read aloud a story during exacerbation (iii) and eleven recordings in which the patient read aloud a story during stable COPD (iv). The sampling frequency of all COPD recordings was 16 kHz.

The control items were selected from the CGN to match the speech of the COPD group. The five controls read a small part of the text 'Papa en Marloes' (Van de Weijer and Slis, 1991). The sampling frequency of the 'healthy' recordings was 16 kHz.

#### 3.3 Procedure

All participants willing to participate received an oral explanation of the research by the data collector, before they consented to participate. The participants were then asked to sustain the vowel [a:] for as long as they could. The data collector marked the beginning of the task by clapping his hands. After the recordings, the patients could take a short break, but no patient took a break for longer than 30 seconds. The participants started reading the assigned part of the story when they were ready. The data collector marked the end of the part by clapping his hands or by verbally telling them to stop. This procedure was repeated on the day the patients were discharged from the hospital, but the participants then read a different part of the story 'De Koning' (Bomans, 1946) to avoid a learning effect.

Each data file was then split in two parts in Praat (Version 6.0.54, Boersma and Weenink, 2019) to separate the sustained vowel from the story. The storytelling recordings were automatically aligned on word and phoneme level using a forced aligner that we developed (Ganzeboom et al., 2016). This alignment was then checked manually by one of the authors in Praat. Subsequently, the aligned files from both the COPD patients and the healthy controls were manually annotated by the same author, using six tiers for the story's transcript, word segmentations, phonetic segmentations, respiratory remarks, speaker noises and commentary.

Based on the annotations and summary per recording, several parameters and measures were manually calculated. The number of produced syllables was manually calculated using the transcript. The number of breath groups (sequence of words, articulated in one single exhalation, before the speaker pauses for breath) as well as the number of inhalations were calculated using the annotation tier. Each inhalation was then compared with a protocol for linguistically acceptable inhalations, based on a set of rules. The number of non-linguistic inhalations consisted of those inhalations that violated these rules. The information on breathing has been obtained manually in the current study. Nallanthighal, Härmä and Strik (2019) showed that it is possible to automatically obtain breathing information from the speech signal. In the future, we intend to use similar algorithms in our COPD research.

Then, a Praat script by Kerhoff (2015) was used to calculate the four formants, total duration, mean frequency, mean intensity, pitch variability, mean center of gravity, total duration of voiced intervals and the total duration of silence

intervals for the recordings containing the story. The files containing the sustained vowels were analysed using Praat’s voice reports (Boersma and Weenink, 2019) to determine several acoustic parameters, namely shimmer (local), shimmer apq3, shimmer apq5, jitter (local), jitter ppq5, jitter rap, harmonics-to-noise ratio (HNR) and the degree of voice breaks. Subsequently, the files containing the sustained vowels were analysed using the script by Kerkhoff (2015) to calculate the four formants, mean frequency, mean intensity and duration. Finally, additional variables were calculated based on the measures that were already obtained: the number of syllables per breath group, speaking and articulation rate, the number of inhalations per breath group, the number of non-linguistic inhalations per breath group and the ratio between voiced and silence intervals (xiii). The Praat measures were not adjusted for possible errors.

### 3.4 Design and Data Analysis

To investigate to what extent various acoustic measures differed for stable COPD patients and healthy controls and for COPD patients in exacerbation and in a stable condition, several one-way analyses of variance (ANOVAs) were conducted. The data were analysed using the statistical software package ‘SPSS’ (SPSS, 2016).

Parameter	Description
Syllables per breath group	The number of syllables produced on one breath
Speaking rate	Speech tempo (in syllables per second) including pauses
Articulation rate	Speech tempo (in syllables per second) excluding pauses
Mean frequency	Perceived pitch in Hz
Mean intensity	Perceived loudness in dB
Pitch variability	Range in variation in level and extent of pitch in Hz
Mean center of gravity	Weighted mean frequency in Hz
Inhalations	Inhalations per syllable
Non-linguistic inhalations	Inhalations in non-linguistic places per syllable
Ratio voiced/silence intervals	Ratio between voiced intervals and silence intervals (-)

Table 1. Overview of acoustic measures used for analysing the sustained vowels and storytelling recordings

[1] *Stable COPD vs. healthy*. First, we compared stable COPD patients with healthy controls (between-subjects design) with respect to ten acoustic measures based on the storytelling recordings. These measures (presented in Table 1) served as the dependent variables of the ANOVAs. Group (stable COPD patients vs. healthy controls) was the independent variable in the analysis (see section 4.1).

[2] *COPD: stable vs. exacerbation*. Subsequently, we conducted two one-way repeated measures ANOVAs to compare the COPD patients in exacerbation and after exacerbation (stable) (within-subjects design).

[2.1] *COPD: stable vs. exacerbation – stories*. In the first analysis, we used the same ten acoustic measures based on the storytelling recordings as dependent variables. Condition (exacerbation vs. stable) was the independent variable (see section 4.2.1).

[2.2] *COPD: stable vs. exacerbation – sustained vowels*. A second one-way repeated measures ANOVA was carried out to assess differences between acoustic measures obtained from recordings of sustained vowels (see Table 2). These acoustic measures served as the dependent variables, and the variable Condition (exacerbation vs. stable) was the independent variable (within-subjects) (see section 4.2.2).

Parameter	Description
Formants	F1, F2, F3 and F4 in Hz
Mean frequency	Perceived pitch in Hz
Mean intensity	Perceived loudness in dB
Duration	Duration (maximum phonation time) in seconds
Shimmer	Variability of the peak-to-peak amplitude in %
Shim apq3	Three-point amplitude perturbation quotient in %
Shim apq5	Five-point amplitude perturbation quotient in %
Jitter	Vocal fold frequency variability from cycle to cycle (frequency perturbation) in %
Jitter ppq5	Five-point perturbation quotient in %
Jitter rap	Relative average perturbation (rap) in %
HNR	Proportion of the harmonic sound to noise in the voice (-)
Voice breaks	Fraction of pitch frames that are analysed as unvoiced in %

Table 2. Overview of acoustic measures used for analysing the recordings of the sustained vowels.

## 4. Results

### 4.1 Stable COPD Patients vs Healthy Controls

Two COPD patients had to be excluded from the analyses, due to inaccurate task execution while reading the story in stable condition, and a damaged file. In total, the analyses were performed on the storytelling recordings of nine COPD patients in stable condition and five healthy controls.

Table 3 shows the mean values and standard deviations of the acoustic measures based on the storytelling recordings split by Group (stable COPD vs. healthy). Moreover, it presents the results of the one-way ANOVA for each of the measures.

Parameter	M (SD)		F	$\eta_p^2$
	Stable	Healthy		
Syllables per breath group	12.78 (3.18)	17.43 (5.46)	4.160 ■	0.26
Speaking rate (syl/sec)	3.67 (0.19)	3.51 (0.25)	0.284	0.02
Articulation rate (syl/sec)	4.64 (0.21)	5.28 (0.28)	3.418 ■	0.22
Inhalations per syllable	0.082 (0.017)	0.056 (0.014)	7.990 *	0.40
Non-ling. inhalations per syllable	0.022 (0.018)	0.00 (0.00)	7.390 *	0.38
Ratio voiced/silence intervals	4.51 (0.66)	2.07 (0.89)	4.876 *	0.29
Mean frequency (Hz)	190.77 (91.69)	128.90 (20.13)	2.145	0.15
Mean intensity (dB)	63.32 (8.12)	67.72 (5.31)	1.161	0.09
Pitch variability (Hz)	710.27 (408.30)	394.74 (116.01)	1.692	0.12
Mean center of gravity (Hz)	504.34 (203.60)	723.37 (255.97)	3.117	0.21

Note: ■  $p < 0.1$ ; \*  $p < 0.05$

Table 3: Overview of the means, standard deviations, and results of the ANOVA (stable vs. healthy) for all variables.

The total number of inhalations per syllable (1) and the number of non-linguistic inhalations per syllable (2) were both significantly higher for COPD patients ( $M_1 = 0.82$ ,  $SD_1 = 0.017$ ;  $M_2 = 0.22$ ,  $SD_2 = 0.018$ ) than for healthy controls ( $M_1 = 0.56$ ,  $SD_1 = 0.014$ ;  $M_2 = 0.00$ ,  $SD_2 = 0.00$ ),  $F(1,12) = 7.990$ ,  $p < 0.05$ ,  $\eta_p^2 = 0.40$  and  $F(1,12) = 7.390$ ,  $p < 0.05$ ,  $\eta_p^2 = 0.38$  respectively. In addition, the ANOVA revealed that the ratio between voiced intervals and silenced intervals was higher for COPD patients ( $M = 4.51$ ,  $SD = 0.66$ ) than for healthy controls ( $M = 2.07$ ,  $SD = 0.89$ ),  $F(1,12) = 4.876$ ,  $p < 0.05$ ,  $\eta_p^2 = 0.29$ . In addition, the effect of Group was marginally significant for the number of syllables per breath group and articulation rate, resp.  $F(1,12) = 4.160$ ,  $p = 0.04$ ,  $\eta_p^2 = 0.26$  and  $F(1,12) = 3.418$ ,  $p = 0.07$ ,  $\eta_p^2 = 0.22$ . The number of syllables per breath group (1) and the articulation rate (2) were lower for COPD

patients ( $M_1 = 12.78$ ,  $SD_1 = 3.18$ ;  $M_2 = 4.64$ ,  $SD_2 = 0.21$ ) than for healthy controls ( $M_1 = 17.43$ ,  $SD_1 = 5.46$ ;  $M_2 = 5.28$ ,  $SD_2 = 0.28$ ). No significant differences between the two groups were found for the other measures ( $p > 0.05$ ).

### 4.2 COPD Patients: Exacerbation vs Stable

#### 4.2.1 Story

For these within-subject analyses, the same two participants were excluded as in the between-subject analyses, because of the same reasons. Therefore, the analyses were conducted on the storytelling recordings of nine COPD patients collected during exacerbation and in stable condition.

Parameter	M (SD)		F	$\eta_p^2$
	Exa.	Stable		
Syllables per breath group	9.50 (1.70)	12.78 (3.18)	12.807 **	0.62
Speaking rate (syl/sec)	3.55 (0.53)	3.67 (0.19)	0.678	0.08
Articulation rate (syl/sec)	5.01 (0.51)	4.64 (0.21)	2.629	0.25
Inhalations per syllable	0.108 (0.021)	0.082 (0.017)	13.545 **	0.63
Non-ling. inhalations per syllable	0.023 (0.021)	0.022 (0.019)	0.014	0.00
Ratio voiced/silence intervals	2.92 (1.83)	4.51 (0.66)	1.844	0.19
Mean frequency (Hz)	154.05 (31.01)	190.77 (91.69)	1.198	0.13
Mean intensity (dB)	64.53 (6.73)	63.32 (8.12)	0.163	0.02
Pitch variability (Hz)	586.53 (238.99)	710.27 (408.30)	1.191	0.13
Mean center of gravity (Hz)	439.12 (152.32)	504.34 (203.60)	1.934	0.20
F1 (Hz)	529.08 (54.08)	555.49 (59.94)	2.089	0.21
F2 (Hz)	1701.46 (112.70)	1731.42 (91.99)	3.137	0.28
F3 (Hz)	2859.84 (152.27)	2871.88 (167.57)	0.282	0.03
F4 (Hz)	3969.85 (201.77)	3956.87 (167.57)	0.091	0.01

Note: \*  $p < 0.05$ ; \*\*  $p < 0.01$

Table 4: Overview of the means, standard deviations, and results of the ANOVA (in exacerbation vs. stable) for all variables.

Multiple one-way repeated measures ANOVAs were carried out to measure the effect of Condition (in exacerbation versus stable COPD) on the dependent variables presented in Table 1. Table 4 provides an

overview of the means, standard deviations, and the results of the ANOVA for all variables.

The analysis revealed a large sized effect of Condition on the number of syllables per breath group  $F(1, 8) = 12.807$ ,  $p < 0.01$ ,  $\eta_p^2 = 0.62$ . The number of syllables per breath group was lower during exacerbation ( $M = 9.50$ ,  $SD = 1.70$ ) than in stable condition ( $M = 12.78$ ,  $SD = 3.18$ ). In addition, the number of inhalations per syllable was significantly higher during exacerbation ( $M = 0.108$ ,  $SD = 0.021$ ) than in stable condition ( $M = 0.082$ ,  $SD = 0.017$ ),  $F(1,8) = 13.545$ ,  $p < 0.01$ ,  $\eta_p^2 = 0.63$ . Other variables did not significantly differ as a function of Condition ( $p > 0.05$ )

#### 4.2.2 Sustained Vowel

For the analyses based on the sustained vowel recordings, we had to exclude two participants. One participant was excluded due to outliers and for the other one we were not able to calculate the acoustic measures. In total, recordings of the sustained vowel by nine COPD patients were included in the analyses.

Parameter	M (SD)		F		$\eta_p^2$
	Exa.	Stable			
Intensity (dB)	65.41 (7.11)	68.77 (7.11)	1.393		0.15
Frequency (Hz)	140.51 (29.48)	150.61 (43.11)	0.792		0.09
Duration (s)	6.41 (2.77)	8.70 (2.52)	37.843	***	0.83
F1 (Hz)	594.12 (119.45)	613.39 (94.37)	0.227		0.03
F2 (Hz)	1327.45 (170.83)	1272.69 (167.82)	1.497		0.16
F3 (Hz)	2829.28 (485.41)	2874.41 (307.11)	0.337		0.04
F4 (Hz)	4070.99 (166.53)	4025.94 (217.49)	0.271		0.03
Jitter (%)	1.32 (0.92)	0.81 (0.45)	2.181		0.22
Jitter ppq5 (%)	0.81 (0.66)	0.45 (0.25)	2.346		0.23
Jitter rap (%)	0.70 (0.48)	0.50 (0.31)	0.914		0.10
Shimmer (%)	8.38 (5.95)	4.18 (1.41)	5.693	*	0.42
Shimmer apq3 (%)	4.23 (3.06)	2.19 (0.92)	4.847	■	0.38
Shimmer apq5 (%)	4.23 (3.47)	2.48 (0.81)	6.244	*	0.44
HNR (-)	15.62 (5.18)	19.49 (2.68)	3.078		0.28
Voice breaks (%)	1.25 (1.81)	0.00 (0.00)	4.237	■	0.35

Note: ■  $p < 0.1$ ; \*  $p < 0.05$ ; \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$

Table 5: Overview of the means, standard deviations, and results of the ANOVA (in exacerbation vs. stable) for all variables.

In order to assess the effect of Condition (exacerbation vs. stable) on the acoustic measures based on the vowel

recordings (see Table 2), we performed multiple one-way repeated measures ANOVAs. Table 5 presents an overview of means and standard deviations by Condition and includes the results of the analysis.

The analysis showed a large effect of Condition on the mean duration of the sustained vowel,  $F(1,8) = 37.843$ ,  $p < 0.001$ ,  $\eta_p^2 = 0.83$ . The mean duration of the sustained vowel was shorter during exacerbation ( $M = 6.41$ ,  $SD = 2.77$ ) than in stable condition ( $M = 8.70$ ,  $SD = 2.52$ ). In addition, the shimmer measures shimmer (1) and shimmer apq5 (2) turned out to be significantly higher in exacerbation ( $M_1 = 8.38$ ,  $SD_1 = 5.95$ ;  $M_2 = 4.23$ ,  $SD_2 = 3.47$ ) than in stable condition ( $M_1 = 4.18$ ,  $SD_1 = 1.41$ ;  $M_2 = 2.48$ ,  $SD_2 = 0.81$ ), resp.  $F(1,8) = 5.693$ ,  $p < 0.05$ ,  $\eta_p^2 = 0.42$  and  $F(1,8) = 6.244$ ,  $p < 0.05$ ,  $\eta_p^2 = 0.44$ . The other shimmer measure, shimmer apq3, was marginally significant ( $F(1,8) = 4.847$ ,  $p = 0.059$ ,  $\eta_p^2 = 0.38$ ) with also higher values in exacerbation ( $M = 4.23$ ,  $SD = 3.06$ ) compared to the stable condition ( $M = 2.19$ ,  $SD = 0.92$ ). Moreover, the analysis revealed a marginally significant effect of Condition on the degree of voice breaks. The degree of voice breaks was higher during exacerbation ( $M = 1.25$ ,  $SD = 1.81$ ) than in stable condition ( $M = 0.00$ ,  $SD = 0.00$ ),  $F(1, 8) = 4.237$ ,  $p = 0.074$ ,  $\eta_p^2 = 0.35$ . The analysis did not show an effect of condition on intensity, frequency, formants, jitter measurements and HNR ( $p > 0.05$ ).

## 5. Discussion

The present research aimed to identify acoustic speech characteristics in both read speech and sustained vowels that mark the beginning of an exacerbation in COPD patients.

[1] *Stable COPD vs. healthy*. The recordings of nine stable COPD patients reading aloud part of De Koning (Bomans, 1946) were compared with the recordings of five healthy controls reading aloud part of Papa en Marloes (Van de Weijer and Slis, 1991). The results showed a significant effect of condition on the number of inhalations per syllable, the number of non-linguistic inhalations per syllable and the ratio of voiced and silence intervals. The number of inhalations per syllable and the number of non-linguistic inhalations per syllable were higher for COPD patients than for healthy controls, which was in line with our expectations based on previous research (GOLD, 2019). However, the higher ratio of voiced and silence intervals for COPD patients compared to healthy controls did not confirm our expectations based on previous research regarding asthmatic patients (Wiechern, Liberty and Pattemore, 2018). There was a trend for the effect of condition on the number of syllables per breath group. The number of syllables per breath group was higher for healthy controls compared to COPD patients, which corresponded with our expectations based on GOLD (2019). There was no effect of condition on pitch, intensity, center of gravity, pitch variability, speaking rate and articulation rate. This was not in line with results from previous studies regarding the differences between pathological and healthy speech (e.g. Alcock et al., 2000; Harel, Cannizzaro and Snyder, 2004; Prelock and Hutchins, 2018; Van Son and Pols, 1996)

[2.1] *COPD: stable vs. exacerbation – stories.* Next, the recordings of nine COPD patients reading aloud a story during exacerbation were compared with the corresponding recordings obtained in stable condition to determine which acoustic measures extracted from read speech differed in the two conditions. The results showed that there was a significant effect of condition on the number of syllables per breath group and the number of inhalations per syllable. The lower number of syllables per breath group and the higher number of inhalations per syllable for COPD patients in exacerbation compared to stable COPD patients were in line with our expectations based on GOLD (2019). There was no effect of condition on pitch, intensity, formants, mean center of gravity, pitch variability, number of non-linguistic inhalations per syllable, ratio between voiced and silence intervals, speaking rate and articulation rate. These results are not in line with those of previous research regarding asthmatic patients or pathological speakers (e.g. Lotan et al., 2019; Shrivastava, Tripathi and Singh, 2018; Wiechern et al., 2018).

[2.2] *COPD: stable vs. exacerbation – sustained vowels.* Finally, the pre-recorded sustained vowels of nine COPD patients during exacerbation were compared with recordings of the same material obtained in stable condition to determine which acoustic measures extracted from sustained vowels differed in the two conditions. The results showed that there was a significant effect of condition on duration, shimmer measurements, HNR and the degree of voice breaks, whereas there was no effect of condition on pitch, intensity, formant frequencies and jitter measurements. The duration was shorter, the HNR was lower and the degree of voice breaks and shimmer measurements were higher for COPD patients in exacerbation compared to stable COPD patients, which confirmed the expectations based on previous research regarding acoustic measures for the identification or monitoring of pathological speech (e.g. Awan, 2011; Boersma, 2004; John et al., 2018; Parsa and Jamieson, 2001; Pinto, Crespo and Mourão, 2014). However, the results regarding pitch, intensity, formants and jitter did not correspond with results from previous studies on asthmatic patients or pathological speakers (e.g. Shrivastava et al., 2018; Teixeira, Oliveira and Lopes, 2013).

For the current research, we made use of pre-recorded data. Unfortunately, there was limited information regarding the participants, their characteristics and methods used to determine the diagnosis. Therefore, the patients could have suffered from comorbidity. The conditions were referred to as ‘in exacerbation’ and ‘stable’. The exacerbations were diagnosed by a medical professional. However, it is possible that not all patients had fully recovered on the day they were discharged. Patients might have preferred a recovery at home, leaving the hospital prematurely. There was great variation in the number of days patients stayed at the hospital (range: 2 – 23 days). This might have resulted in an insufficient contrast between the in exacerbation condition and the stable condition.

The results are promising and different measures, such as HNR, shimmer, duration, the number of syllables per breath group and the number of inhalations per syllable, show potential for the non-invasive, remote monitoring of COPD patients. The small sample sizes and the limited information available for the COPD patients might have

contributed to the relatively large standard deviations. Larger sample sizes and more information on the patients, could yield smaller standard deviations and more significant results. For instance, then patients could be grouped based on etiology, age, years since diagnosis or the severity of the COPD (see e.g. Gupta et al., 2016). After all, the goal is to detect changes within a COPD patient, and now we are analysing collapsed data of different COPD patients.

Tehrany, Barney and Bruton (2014) mentioned that spontaneous speech is more valuable and suitable for the detection of speech abnormalities in pathological speakers. The tasks used in this study consisted of reading aloud a phonetically balanced text and sustaining a vowel. Such speech tasks have the advantage that they are more controlled, and thus offers more opportunities for analysis and comparisons between different patients and different occasions. A possible scenario for this technology might be as follows. A COPD patient installs an app on his or her mobile phone and gives permission to monitor his or her speech. On the background, all spontaneous speech is monitored and checked for markers of an upcoming exacerbation. When such markers are present, the user is asked to carry out a few controlled tasks, such as sustaining a vowel and reading part of a story. An analysis of these controlled speech tasks then makes better comparison and diagnosis possible. To study whether such a scenario is feasible, for both COPD patients and other patient populations, appropriate speech data should be recorded in realistic conditions and next be analysed carefully. Ideally, COPD patients using a pilot version of such an app could agree to share the recordings obtained in pre-exacerbation, so that these could be analysed to better approach the exacerbation condition. In any case, the current exploratory study already shows promising results, so in the future this might be a feasible scenario.

## 6. Conclusions

The outcomes of this research have provided insight into the acoustic differences between the speech of COPD patients during exacerbation, COPD patients in stable condition and healthy controls.

This research has shown that the speech of COPD patients in exacerbation differs from their speech in stable condition. The HNR, shimmer measurements and duration of a sustained vowel might have potential for the detection of exacerbations, as well as the number of syllables per breath group and the number of inhalations per syllable. However, sustained vowels rarely occur in spontaneous speech. Therefore, the last two outcome measures might have greater potential for the detection of beginning exacerbations. After the initial warning based on the number of syllables per breath group and the number of inhalations per syllable, the patients could be presented with a quick check in the form of a sustained vowel. After production of the sustained vowel, HNR, shimmer and duration could be calculated to reject or confirm the initial warning. The results of the present study are promising, but it is clear that further research into the different outcome measures and their potential is still needed in order to develop applications that are capable of detecting exacerbations in the speech of COPD patients.

## 7. Bibliographical References

- Alcock, K. J., Passingham, R. E., Watkins, K., and Vargha-Khadem, F. (2000). Pitch and timing abilities in inherited speech and language impairment. *Brain and Language*, 75(1):34–46.
- American Speech Language Hearing Association (ASHA). (2014). Telepractice: overview. Retrieved June 2, 2019, from: <https://www.asha.org/Practice-Portal/Professional-Issues/Telepractice/>
- Awan, S. N. (2011). The effect of smoking on the Dysphonia Severity Index in females. *Folia Phoniatrica et Logopaedica*, 63:65–71.
- Boersma, P. (2004). Stemmen meten met Praat. *Stem-Spraak en Taalpathologie*, 12(4):237–251.
- Boersma, P., and Weenink, D. (2019). Praat: doing phonetics by computer [Computer program]. Version 6.0.54. Retrieved June 1, 2019, from: <http://www.praat.org/>
- Bolton, C. E., Waters, C. S., Peirce, S., and Elwyn, G. (2010). Insufficient evidence of benefit: a systematic review of home telemonitoring for COPD. *Journal of Evaluation in Clinical Practice*, 17(6):1216–1222.
- Bomans, G. (1946). *Sprookjes*. Amsterdam: Elsevier.
- Boyanov, B., and Hadjitodorov, S. (1997). Acoustic analysis of pathological voices. A voice analysis system for the screening of laryngeal diseases. *IEEE Engineering in Medicine and Biology*, 16(4):74–82.
- Buist, A. S., McBurnie, M. A., Vollmer, W. M., Gillespie, S., Burney, P., Mannino, D. M., ... Nizankowska-Mogilnicka, E. (2007). International variation in the prevalence of COPD (The BOLD study): a population-based prevalence study. *Lancet*, 370:741–750.
- Chen, J., Zhu, J., and Ye, J. (2019). An attention-based hybrid network for automatic detection of Alzheimer’s Disease from narrative speech. *Proceedings of Interspeech 2019*, Graz, Austria, 4085 – 4090.
- Constantinescu, G., Theodoros, D., Russell, T., Ward, E., Wilson, S., and Wootton, R. (2010). Assessing disordered speech and voice in Parkinson’s disease: a telerehabilitation application. *International Journal of Language & Communication Disorders*, 45(6):630–644.
- Ganzeboom, M., Bakker, M., Cucchiari, C., Strik, H. (2016). Intelligibility of Disordered Speech: Global and Detailed Scores. *Proceedings of Interspeech 2016*, San Francisco, USA, 2016.
- GOLD: Global Initiative for Chronic Obstructive Lung Disease. (2019). Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease [Report]. Accessed June 2, 2019, from: <https://goldcopd.org/gold-reports/>
- Gupta, N., Pinto, L., Benedetti, A., Li, P. Z., Tan, W. C., Aaron, S. D., ... Bourbeau, J. (2016). The COPD Assessment Test: can it discriminate across COPD subpopulations? *Chest*, 150(5):1069–1079.
- Gutz, S. E., Wang, J., Yunusova, Y., and Green, J. R. (2019). Early identification of speech changes due to amyotrophic lateral sclerosis using machine classification. *Proceedings of Interspeech 2019*, Graz, Austria, 604 – 608.
- Haasnoot, R. (2012). De ontwikkeling van de fonetisch uitgebalanceerde standaardtekst De Koning [Master thesis]. Radboud University. Retrieved June 2, 2019, from: <http://haasnoot.net/sitebuildercontent/sitebuilderfiles/dekoningrhaasnootdefinitief.pdf>
- Harel, B., Cannizzaro, M., and Snyder, P. J. (2004). Variability in fundamental frequency during speech in prodromal and incipient Parkinson’s disease: a longitudinal case study. *Brain and Cognition*, 56(1):24–29.
- Hurst, J. R., Vestbo, J., Anzueto, A., Locantore, N., Müllerova, H., Tal-Singer, R., ... Wedzicha, J. A. (2010). Susceptibility to exacerbation in chronic obstructive pulmonary disease. *The New England journal of medicine*, 363(12):1128–1138.
- John, K., Balagangu, S., Jeromy, A. K., Albert, D. A., Christopher, N., and Godfred, B. A. (2018). Speech signal analysis as an alternative to spirometry in asthma diagnosis. *Journal of Biomedical Engineering and Medical Devices*, 3(3):1–6.
- Klersy, C., De Silverstri, A., Gabutti, G., Raisaro, A., Curti, M., Regoli, F., and Auricchio, A. (2014). Economic impact of remote patient monitoring: an integrated economic model derived from a meta-analysis of randomized controlled trials in heart failure. *European Journal of Heart Failure*, 13(4):450–459.
- López, C. C., Serdá, G. J., Lacalzada, C. C., Medina, A. M., Blanco, J. A. G., Bello, M. A. G., and Navarro, P. C. (2014). Prevalence of Chronic Obstructive Pulmonary Disease in the Canary Islands. *Archivos de Bronconeumologia*, 50(7):272–277.
- Lotan, C., Kremer-Tal, S., Lotan, A., Schlik, Z., Gerner, A., Snir, Y., ... Sheinkerman, M. (2019). Assessment of a pulmonary condition by speech analysis [patent application]. US Patent Application Publication No. US 2019/0080803 A1.
- Maryn, Y., and Roy, N. (2012). Sustained vowels and continuous speech in the auditory-perceptual evaluation of dysphonia severity. *Jornal da Sociedade Brasileira de Fonoaudiologia*, 24(2):107 – 112.
- Nallanthighal, V. S., Härmä, A., Strik, H. (2019) Deep Sensing of Breathing Signal During Conversational Speech. *Proceedings of Interspeech 2019*, Graz, Austria, 4110 – 4114.
- Ong, M. K., Romano, P. S., Edgington, S., Aronow, H. U., Auerback, A. D., Black, J. T., ... Fonarow, G. C. (2016). Effectiveness of remote patient monitoring after discharge of hospitalized patients with heart failure. *JAMA Internal Medicine*, 176(3):310–318.
- Oostdijk, N. (2000). Digitaal. Het Corpus Gesproken Nederlands. *Nederlandse Taalkunde*, 5(3):280–285.
- Parsa, V. and Jamieson, D. G. (2001). Acoustic discrimination of pathological voices: sustained vowels versus continuous speech. *Journal of Speech, Language and Hearing Research*, 44(2):327–339.
- Pinto, A. G. L., Crespo, A. N., and Mourão, L. F. (2014). Influence of smoking isolated and associated to multifactorial aspects in vocal acoustic parameters. *Brazilian Journal of Otorhinolaryngology*, 80(1):60–67.
- Postma, D. S., Bush, A., and Van den Berge, M. (2015). Risk factors and early origins of chronic obstructive pulmonary disease. *Lancet*, 385(9971):7–13.

- Prelock, P. A. and Hutchins, T.L. (2018). Children with speech disorders. In P. A. Prelock and T. L. Hutchins, *Clinical guide to assessment and treatment of communication disorders. Best practices in child and adolescent behavioural health care*. Cham, Switzerland: Springer, pp. 75–87.
- Robinson, D. J., Jerrard-Dunne, P., Greene, Z., Lawson, S., Lane, S., and O’Neill, D. O. (2011). Oropharyngeal dysphagia in exacerbations of chronic obstructive pulmonary disease. *European Geriatric Medicine*, 2(4):201–203.
- Rutowski, T. Harati, A., Lu, Y., and Shriberg, E. (2019). Optimizing speech-input length for speaker-independent depression classification. *Proceedings of Interspeech 2019*, Graz, Austria, 3023 – 3027.
- Shor, J., Emanuel, D., Lang, O., Tuval, O., Brenner M., Cattiau, J., ... Matias, Y. (2019). Personalizing ASR for dysarthric and accented speech with limited data. *Proceedings of Interspeech 2019*, Graz, Austria, 784 – 788.
- Shrivastava, P., Tripathi, N., Singh, B. K. (2018). Assessment of respiratory disorders using speech parameters. *International Journal on Future Revolution in Computer Science & Communication Engineering*, 4(2):461–464.
- Soriano, J. B., Zielinski, J., and Price, D. (2009). Screening for and early detection of chronic obstructive pulmonary disease. *Lancet*, 374:721–732.
- SPSS (2016) IBM Corp. Released 2016. IBM SPSS Statistics for Windows, Version 24.0. Armonk, NY: IBM Corp.
- Tehrany, R., Barney, A., and Bruton, A. Differentiating between health and respiratory disease using speech breathing pattern analysis. *European Respiratory Journal*, 23(44):58.
- Teixeira, J. P., Oliveira, C., and Lopes, C. (2013). Vocal Acoustic Analysis - jitter, shimmer and HNR parameters. *Procedia Technology*, 9:1112–1122.
- Trappenburg, J. C., Monnikhof, E. M., Bourbeau, J., Troosters, T., Schrijvers, A. J., Verheij, T. J., and Lammers, J. W. (2011). Effect of an action plan with ongoing support by a case manager on exacerbation-related outcome in patient with COPD: a multicentre randomised controlled trial. *Thorax*, 66(11):977–984.
- Van de Weijer, J. C., and Slis, I. H. (1991). Nasaliteitsmeting met de nasometer. *Logopedie en Foniatrie*, 63:97–101.
- Van Son, R. J. J. H., and Pols. L. W. (1996). An acoustic profile of consonant reduction. *Conference: the 4th International Conference on Spoken Language Processing*, Philadelphia, PA, USA, 1529 – 1532.
- Wiechern, K. A., Liberty, P., and Pattemore, E. (2018). The effects of asthma on breathing during reading aloud. *Speech, Language and Hearing*, 21(1):30–40.
- Vyshedskiy, A., and Murphy, R. (2016). Acoustic biomarkers of chronic obstructive lung disease. *Research Ideas and Outcomes*, 2:e9173.

## 8. Acknowledgements

We would like to express our gratitude to Hanneke van Helvoort from the department of Pulmonary Diseases of the Radboud University Medical Center for her contribution to the data collection for this study. We would also like to thank Wei Xue for her supervision during the aligning process. Finally, we want to acknowledge the support of Mario Ganzeboom, Joop Kerkhoff and Maarten Vos, for providing us with scripts to align and analyse our data.



# Latent Feature Generation with Adversarial Learning for Aphasia Classification

Anna Vechkaeva, Günter Neumann

German Research Center for Artificial Intelligence (DFKI)

Saarbrücken, Germany

{anna.vechkaeva, guenter.neumann}@dfki.de

## Abstract

Aphasia is a language disorder resulting from brain damage, and can be categorised into types according to the symptoms. Automatic aphasia classification would allow for quick preliminary assessment of the patients' language disorder. A supervised approach to automatic aphasia classification would require substantial amount of training data, however, aphasia data is sparse. In this work, we attempt to use data generation, namely Generative Adversarial Networks (GANs), to deal with data sparsity. The latent feature generation approach is used to deal with the text generation non-differentiability problem, which is an issue for GANs. The approach using artificially generated data to augment training set was tested. We conclude through running a series of experiments that it has potential to improve aphasia classification in the context of low resource data, provided that the available data is enough for the generative model to properly learn the distribution.

**Keywords:** Adversarial Learning, Feature Generation, GANs, Aphasia

## 1. Introduction

Aphasia is a language disorder resulting from brain damage, such as stroke, physical damage, or degenerative dementias. Depending on the brain region, which was damaged, and the severity of the damage, it can manifest with various symptoms, which differ from patient to patient. Aphasia can be categorised into different types, which require different kinds of therapy. Therefore, automatic aphasia type classification could be beneficial for aphasia patients as well as speech therapists, as it would allow for quick preliminary assessment of patients' language disorder, and consequently faster therapy selection.

Although Allen et al. (2012) provides evidences that there exists effective therapy for chronic aphasia, multiple studies suggest that recovery after stroke (Kinsella and Ford, 1980; Skilbeck et al., 1983; Demeurisse et al., 1980), as well as after other brain damage (Jennett and Bond, 1975; Bond and Brooks, 1976) mostly happens in the first several months after the incident leading to the brain damage, and very little, if any, progress happens after one year (Hanson et al., 1989). Providing intensive therapy as soon as possible is crucial for rehabilitation and lack of it can compromise the outcome of the patients' recovery (Bhogal et al., 2003).

In this work, we attempt to classify the aphasia types in the context of low resource data. For this, deep neural networks (DNNs), as well as other machine learning algorithms were used. Using DNNs in this problem is challenging because they normally require a big amount of data to train successfully. One of the ways of dealing with data sparsity attempted in this work is generating synthetic data and using this data for training. Unlike (Chen et al., 2019) who generated structured data for patients' medical records, we focus on generating representations of unstructured textual data. We test if generating synthetic data can help improve the classification, focusing on the Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) framework

due to their success in generative modeling. We adopt an adversarial feature learning approach (Ganin et al., 2016), which does not require generating actual textual data. Unlike GANs, which aim to generate realistic data, the adversarial feature learning approach generates the hidden representation of the data. This approach is suitable for non-generative tasks, such as classification, and alleviates the need of generating actual textual data, which notably has limitations with GANs.

In this work, we attempt to develop a model which, given a participant's speech transcript, predicts an aphasia type label. For this, we will first classify each individual phrase produced by the participant and use these labels to predict the participant's overall score. As the aphasia data is sparse we will also use a generative model to augment the training set. For this we will generate utterance level hidden representations, which will later be used for training utterance level classification model. The main contributions of this work are: (1) testing an approach of using GANs in the context of sparse data for aphasia classification, and (2) developing a latent feature generation approach to solve the GANs issues when dealing with text.

## 2. Background

### 2.1. Aphasia Classification

Aphasia was first described by French neuroanatomist Paul Broca (Broca, 1861) and since then, multiple aphasia classifications have been suggested. One of the ways to classify the patients into groups is using one of the standard protocols, for example, Western Aphasia Battery (WAB) (Kertesz, 2007). It distinguishes between the following aphasia types: Broca's, Wernicke's, anomic, conduction, transcortical motor, transcortical sensory, and global. These aphasia types differ by symptoms and severity. This aphasia classification scheme, as well as other ones, has been criticised, because often, patients' symptoms cannot be fit into one type and there exists overlap between the classes (Caramazza, 1984; Swindell et al., 1984). Nevertheless, WAB

provides a way to categorize patients according to their most prominent symptoms. Moreover there are datasets labeled using WAB scheme, which is important for studies using methods requiring substantial training data.

The speech of the people suffering from aphasia of different types and severity has distinctive characteristics, which can help to automatically analyze aphasia. There are studies which describe different features of aphasia speech in comparison to healthy speakers, as well as features specific to different aphasia types. These features can be acoustic (Damasio, 1992; Leung et al., 2017), grammatical (Kolk, 1998), discourse (Ulatowska et al., 1981) and semantic (Jefferies and Lambon Ralph, 2006). The presence of these features in aphasic speech suggests that it is possible to use them to automatically categorize aphasia.

Aphasia classification task is a problem which has been approached by researchers in the past. Järvelin and Juhola (2011) provide a system for distinguishing speech of people with aphasia from healthy controls' speech, and compare different machine learning techniques for identifying aphasia speakers. There are a number of studies, where authors attempt to distinguish different types of aphasia from each other, using groups of features. For example, Yourganov et al. (2015) attempt to predict types of aphasia based on fMRI brain images of the patients. There are studies, which assess aphasia, based on features extracted from other language production modalities like writing (Basso et al., 1978), sign language (Marshall et al., 2004), and comprehension (Mesulam et al., 2015; Purdy et al., 2019).

In order to analyze impaired speech, authors use two different kinds of features: acoustic and textual. For example, Qin et al. (2018) propose a system for assessing aphasia speech using textual features. The aphasia severity is predicted based on syllable level vectors, acquired from text produced by automatic speech recognition system, given recording of aphasia speech. Fraser et al. (2014) extract features from aphasia speech transcripts and use them to classify primary progressive (slow impairment of language caused by neurodegenerative disease) aphasia types. Themistocleous et al. (2018) identify mild cognitive impairment from speech using acoustic features. They predict if the patient has cognitive impairment based on features such as vowel formants, fundamental frequency and vowel duration. In Little et al. (2009) and Meilán et al. (2014) acoustic features were proven useful for detecting Parkinson's and Alzheimer's disease respectively from the patients' speech. Also, there are studies that provide evidence that combining acoustic and textual information helps to identify Alzheimer's disease (Fraser et al., 2016) and mild cognitive disease (Themistocleous et al., 2018). Although language impairments in case of dementias, can differ in their nature from ones in aphasia due to brain damage, similar approaches can be used to identify and assess them.

## 2.2. Synthetic Data Generation for Classification

The generative models are widely used to tackle the data sparsity problem in various fields and there is work on synthetic data generation for improving text analysis. For example, Maqsd (2015) tests different text generation methods for augmenting the available training data with

synthetic samples for sentiment analysis of text. In this work, methods such as Latent Dirichlet Allocation (LDA), Markov Chain (MC), and Hidden Markov Model (HMM) are tested and the authors conclude that the models can generate the data with the features belonging to each class. In computer vision, synthetic data generation is also used to augment the sparse training data. The Generative Adversarial Networks (GANs) are used to improve different classification tasks, as GANs are able to generate realistic images. Frid-Adar et al. (2018) use GANs for generating the additional image data for improving liver lesion classification. In the paper, the authors train a separate generative model for each of the classes and then use the models to generate the data for the respective classes. A significant improvement of the classification after adding the generated data to the training set is reported. GANs have also been used to generate additional data for text classification. Guan et al. (2018) use conditional GANs to generate electronic medical records.

## 2.3. Generative Adversarial Networks

GANs were proposed by Goodfellow et al. (2014) and showed great success in image generation, becoming very popular. Given training data, the model learns the distribution of the data and produces data instances which belong to this distribution. GAN framework is derived from a game theoretic formulation, where each player can be seen as an adversary. GANs consist of two models: Generator and Discriminator. Given a noise (from normal distribution) as an input, generator's goal is to produce data samples which look like they belong to the same distribution as real data. When fed with the real data samples and samples produced by generator, discriminator's goal is to be able to distinguish between real and fake data. The discriminator's loss is then propagated back to generator so that it can improve and generate more realistic synthetic samples. The conditional GANs (cGAN), which were proposed by Mirza and Osindero (2014) are a type of GAN, that can be conditioned on some extra information. It learns to not only produce datapoints which look realistic but also conditions the produced datapoints with additional class information. GAN models are known to have the training stability issues, meaning that the model does not converge. Other problems which may occur when using GANs are a mode collapse problem and the vanishing gradient problem (Goodfellow, 2016). A number of improved training techniques were proposed (Dziugaite et al., 2015; Huszár, 2015; Li et al., 2015; Salimans et al., 2016; Nguyen et al., 2017; Nowozin et al., 2016; Zhao et al., 2016) since the original introduction of GANs (Goodfellow et al., 2014). The most stable and robust version of GANs, called Wasserstein GAN (WGAN), was proposed by Arjovsky et al. (2017). Wasserstein GAN uses a different loss function for the discriminator, which is called critic in this setting. Instead of classifying the generated samples as real or fake, the critic tries to predict how close the produced samples are to the real distribution. Arjovsky et al. (2017) concluded that when using Wasserstein distance, problems such as mode collapse and vanishing gradient did not appear and the training was more stable.

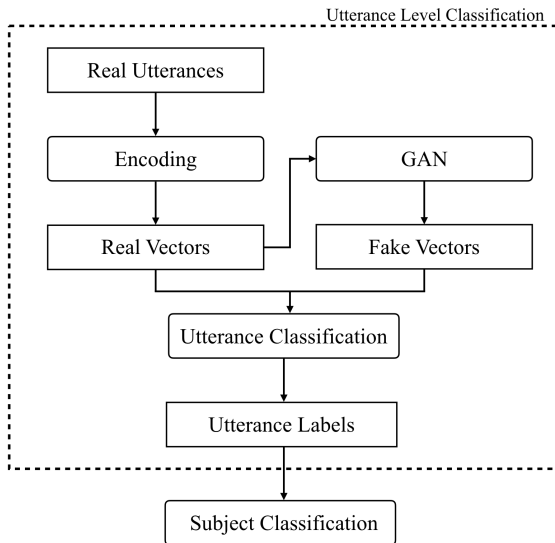


Figure 1: Classification Model with data generation

GANs have one design limitation on the Generator, that it cannot have discrete outputs. This makes them incompatible directly for NLP. By construction, the generative network has to be fully differentiable. Consequently, GAN framework prohibits generator from having discrete outputs. Also, it is not trivial to assign discriminator probabilities to the sequences which are not completely generated (Goodfellow, 2016). The Adversarial Feature Learning (AFL) approach is similar to the GANs approach. While in GANs the adversary aims to determine if the outputs are real or generated, in the AFL, the adversary is created over hidden features. This approach is well suited for non-generative tasks, like a sentence classification task, where the objective is not to classify sentences as real and fake. This approach allows the model to deal with continuous data, which is easier than dealing with the discrete outputs.

### 3. Method

#### 3.1. Overview

For all experiments, the general approach taken in this work of classifying aphasia types is as follows: first, the utterances produced by a subject (or a person) are classified as one of the aphasia types, and after this, based on the utterance level classification, a subject is classified as healthy or having one of the aphasia types. This pipeline includes two classification models: utterance level classifier and subject level classifier. Different versions of both models were also tested in this work.

Figure 1 demonstrates the process of training the model involving artificial data generation. After encoding the real data into vectors, the conditional GAN model is trained using these vectors and corresponding labels. Then, using the trained GAN, the fake data vectors belonging to a specified class given the class label are generated. Following this, an utterance level classifier, trained on both real and generated

data, is used to predict the utterance level labels on the test data, and the subject level classification is run, given the labels produced on the previous step, as an input, to predict the participant level aphasia type labels. The reason for doing utterance level classification first and then subject level classification, instead of doing the subject level classification directly, is that Neural Network approaches normally require a sufficient amount of training data. While the number of subjects in the dataset used in this work is small, the number of utterance level datapoints is much bigger.

#### 3.2. Data

AphasiaBank (MacWhinney et al., 2011) is one of the few publicly available datasets in the aphasia domain. It contains recordings of people suffering from aphasia as well as transcripts of their speech which also provide information about a patient including aphasia type. AphasiaBank also includes interviews with healthy participants recorded following a similar procedure. Table 1 shows an example of a transcript of speech belonging to a patient with anomic aphasia.

---

#### Anomic Aphasia

---

- 1 INV: how do you think your speech is these days ?
  - 2 PAR: uh it's ... it's good but it's very slow .
  - 3 INV: do you remember when you had your stroke ?
  - 4 PAR: um it's two years ago .
  - 5 PAR: and when I when I when I had the stroke  
I couldn't say a word for a year and a half .
- 

Table 1: Example of utterances produced by a patient with Anomic Aphasia (source: AphasiaBank)

The aphasia type labels and aphasia severity scores provided in AphasiaBank are obtained using Western Aphasia Battery (WAB) (Risser and Spreen, 1985). WAB is an instrument for evaluating clinical aspects of language function for individuals with neurological disorders resulting from stroke, brain injury or dementia. It helps to identify presence, severity and type of aphasia and measures linguistic (speech, fluency, auditory comprehension, reading and writing) and non-linguistic performance of individuals. The dataset, which was constructed in this work, consists of utterance level data-points, which are transcriptions of the phrases produced by a subject in response to the interviewer's question. The utterance level datapoints are grouped into subject level datapoints and each of them consists of transcribed utterances produced by one subject. This is done so that the subjects as well as utterance level classification can be performed. AphasiaBank does not provide utterance level aphasia type labels, so the labels for the utterances are assigned based on the aphasia type of the participant who produced the utterance.

Not all utterances, produced by patients suffering from aphasia, contain signs indicating the aphasia type, some of them are completely correct, as shown in the example in Table 1, where utterance 2 is grammatically correct. AphasiaBank provides aphasia type labels only on a subject level, but not on the utterance level. The fact that not all of the

	Utterance level		Subject level	
	Train	Test	Train	Test
Broca’s	2682	1029	66	30
Anomic	10767	2814	108	30
Conduction	4141	3859	34	30
Control	29969	4073	217	30
Total	47559	11775	425	120

Table 2: Number of utterance and subject level datapoints for training and test set

utterances by aphasia patients are aphasic provides a challenge for constructing an utterance level dataset, as aphasia type labels for each utterance cannot be confidently determined. This introduces a certain level of noise to both training and test datasets, as there are a number of non-aphasic utterances marked as aphasic ones. This provides challenges for both model training and the evaluation of results.

The dataset used for training and evaluating our models included the classes, which contained at least 60 patients, so that at least 30 subjects could be used in the test set. The final dataset contained the following classes: Broca’s, Anomic, Conduction and Control. The number of both utterance and subject level datapoints belonging to each class are represented in Table 2.

### 3.3. Hidden Text Representation

In this work, 300 dimensional word vectors pretrained by Google using word2vec (Mikolov et al., 2013) were used for generating hidden text representation<sup>1</sup>. Each utterance from the dataset is represented as a two dimensional matrix constructed from the vectors of each word. In order to make all of the utterance representations have the same dimensions, we add 300 dimensional vectors of zeros to the utterance representations until all representations have the same dimensionality.

### 3.4. Models

#### 3.4.1. Utterance Classification

A Convolutional Neural Network (CNN) (LeCun et al., 1989), which takes a two dimensional vector of an utterance as input and produces probabilities of the utterance belonging to each of the considered classes, was used for utterance classification in all the experiments. The architecture of the model is the same for all the experiments and contains three layers, with the first layer being a linear layer, which flattens the input. This is done so that the vertical relations between the values in word vectors are not taken into account, as, intuitively, unlike in the numerical image representations, there should be no vertical correlation between the individual values of the word vectors. The next layer is a layer with the ReLU activation function.

#### 3.4.2. Subject Classification

The aim of a subject level classification model is to predict a participant’s aphasia label, given utterance level labels predicted for this participant by the utterance level classifier.

In this work, we test two approaches: supervised and unsupervised. The advantage of the unsupervised models is that they do not require additional data to train, therefore the whole training dataset can be used to train the generation and utterance level classification models. On the other hand, we expect that the supervised models will show better results than the unsupervised models, as they will observe the actual distribution of the utterance level labels over the training data. However, as a part of the data will need to be held out during the steps presiding the subject level training, the quality of the generated data can drop.

In an unsupervised approach, a fairly simple model was used. After all the utterances were classified by the utterance level model, the subject was assigned the aphasia type which was most present among the subject’s utterances, according to the classifier. For example, if a subject had 20 utterances classified as Broca’s aphasia, 30 utterances - as Anomic aphasia, 10 utterances - as Conduction aphasia, and 5 sentences - as non-aphasic sentences, the patient was classified as having Anomic aphasia. This algorithm will be further referred to as the Max Class model. Different variations of the described model were also tested. As patients suffering from aphasia are still able to produce non-aphasic sentences, it makes sense to reduce the impact of the non-aphasic utterance on the patient level classification. So, instead of taking into account the entirety of non-aphasic sentences, this number is reduced by dividing it by a range of integers from 2 to 7.

The number of supervised machine learning algorithms were also tested to predict an aphasia type on the subject level, given the utterance level labels predicted by the utterance level classifier. Given a number of utterances for which an utterance level classifier predicted each type of aphasia, it predicted a patient level aphasia type. The algorithms used are Naïve Bayes (NB), Multinomial Naïve Bayes (MNB), Random Forest (RF), Decision Trees (DT), K-Nearest Neighbours (KNN), and Support Vector Machines (SVM).

#### 3.4.3. Generation Model

In this work, we aim to generate hidden features for synthetic data without generating the actual data for text classification using cGANs. We use the static text representations to train the generative model, to produce a hidden representation of the data from the different classes. The model does not take into account other subject properties, like age or gender. That way, the model aims to generate a vector representation of an utterance belonging to a given class without generating actual textual data. The fact that we create the adversary over the hidden features makes this approach similar to the AFS.

For data generation, two types of GANs were tried. The first one is a simple conditional GAN. Both generator and discriminator of this model are CNNs. Binary crossentropy is used as a loss function and the GAN is conditioned on aphasia type and produces vector representation of utterance level datapoints, given a class. The models trained for different amount of epochs were tested. Conditional WGAN, where both generator and critic are CNNs and Wasserstein loss is used, is also tested. It is also condi-

<sup>1</sup><https://github.com/mmihaltz/word2vec-GoogleNews-vectors>

tioned on aphasia type labels. The models trained for the different amount of epochs are also tested. In this work, the Keras implementation of GANs<sup>2</sup> for cGAN was used and for WGAN was used.

### 3.5. Experiments

#### 3.5.1. Baseline

The baseline system uses only real data from AphasiaBank as training data, and no data generation happens at this step. A CNN model trained on the whole training set containing real utterance level datapoints was used to predict an aphasia class for an utterance given its vector representation. For the subject level classification, the Max Class model without any additional alternations was used.

#### 3.5.2. GAN models comparison

The different GAN models that were trained for a different number of epochs were compared, as it is difficult to tell if the GAN converged based only on generator and discriminator losses and sometimes GANs mode collapse can occur, which means that the generator starts producing very similar outputs to trick the discriminator. The conditional GAN trained for 20000 epochs and 5000 epochs and Wasserstein conditional GAN trained for 500 epochs, 1000 epochs, and 2000 epochs models were compared.

In order to assess how good the data produced by each of the tested models is, we investigate how this data can help with the aphasia classification problem. The produced data was used to train an utterance level CNN classifier which was later used to predict aphasia classes for utterances belonging to patients from the test set. After this, these predictions were used by the Max Class model to assign an aphasia type to each patient in the training set.

For each of the models, two different experiments were performed. The first experiment aims to investigate how good the performance of the classification model trained only on generated data is. Each of the GAN models described above was trained on the whole training set to produce utterance level datapoints belonging to a given aphasia class. The generator model produced by each of the models was then used to generate synthetic training data, producing 40000 datapoints (10000 datapoints for each class) by each of the models. Then, an utterance level CNN classifier was trained using the generated data as a training set and max class model was used to predict patient level classes.

The purpose of the second experiment was to assess how combining generated data with the real data can help improve the aphasia type classification. For this experiment, the synthetic data was generated for the aphasia types which have fewer utterance level datapoints in the Aphasia-Bank, so that each class has the same number of datapoints, resulting in 27 287 generated datapoints for Broca’s aphasia, 19 202 - for Anomic aphasia, and 25 828 - for Conduction aphasia. The control group contains the biggest amount of datapoints, so no data was generated for this class. After the data generation step, all the produced datapoints were added to the original training set and each class ended up having 29 969 utterance level datapoints. After this, the CNN utterance level classifier was trained on the

	Utterance level		Subject level	
	Train1	Train2	Train1	Train2
Broca’s	5433	5406	33	33
Anomic	1346	1409	54	54
Conduction	2140	2264	17	17
Control	15115	15309	108	109
Total	24034	24388	212	213

Table 3: Number of utterance and subject level datapoints for two training tests

combined dataset and the Max Class model was used to predict the aphasia type of the subject.

#### 3.5.3. Max Class Model Experiments

As not all of the utterances produced by aphasia patients show signs of aphasia, reducing the impact of the control class on the subject level classification could help to improve the classification. In order to reduce the impact of the control class, we use the number of the control class predictions divided by some number instead of the full number. For example, if the majority of the utterances produced by a subject are classified as non-aphasic, but some other aphasia type class is very present amongst the utterances, the subject might still have aphasia.

The aim of this experiment is to determine by how much the impact of the control class should be reduced. To determine by how much the number of predicted control class should be divided, a range of numbers from 1 to 7 are tested. We assume that initially the classification accuracy will increase as the number becomes bigger, but will start dropping after a point when the impact of the control class will become too small. We also compare the performance of each Max Class model trained on the real data with the one trained on the combination of real and generated data.

#### 3.5.4. Supervised Subject Classification Methods Experiments

In addition to Max Class model a number of machine learning (ML) algorithms were tested for the subject level classification: NB, MNB, RF, DT, KNN, and SVM. Given the labels produced by an utterance level classifier, the classifier should predict the aphasia type of the subject. The Scikit Learn (Pedregosa et al., 2011) implementation of the models listed above was used.

This approach required splitting training data into two parts, as the algorithms used need data to be trained on as well as the utterance level classifier. The utterance level classifier and the subject level ML models need to be trained on different data, because if both of the models will be trained on the same data, the CNN classifier will have to make predictions, which will later be used for training by the ML algorithms for the samples it observed during the training. Our concern is that, given that the data is noisy, the prediction quality for this data will be too different from the predictions made for unseen data. Similarly, the GAN model should not be trained on the data, which later will be used for the ML models training, because if the generative model produces data similar to the data which will be later used for the subject level model training, the model will still

<sup>2</sup><https://github.com/eriklindernoren/Keras-GAN>

Model	Utt. level	Subj. level
Baseline	0.45	0.44
GAN 5000 epochs	0.46	0.39
GAN 20000 epochs	0.46	0.43
W-GAN 500 epochs	0.46	0.42
W-GAN 1000 epochs	0.45	0.48
W-GAN 2000 epochs	0.46	0.45

Table 4: Classification accuracy for GAN models trained for different amount of epochs (trained on both real and generated data)

be indirectly trained on this data through the GAN, and the predictions on this data will be different from the predictions on the unseen data. Therefore, the data was divided subject-wise into two equal parts, so that each part contains the same number of patients per aphasia type class. The number of utterance and subject level datapoints for each class is presented in Table 3. The first half of the data contains 212 subjects and 24 034 utterances, while the second half of the data contains 213 subjects and 24 388 utterances. To compare the performance of the subject level classification models with the approach used before, we also run the Max Class models on the newly divided data. We used only the first part of the training data to train CNN classifier, while the second part of the data was not used, as Max Class models do not need training.

## 4. Results

### 4.1. Evaluation

To assess the performance of the models, accuracy and F1 score are used. On the utterance level only the accuracy is reported. The performance of the models, evaluated on the utterance level test set, does not really reflect the quality of the models. The reason for this is that the test set contains noisy data, because the subject level labels are used to assign labels to utterances when constructing the test set. Because the dataset has gold-standard labels on the subject level, unlike the utterance level evaluation, the subject level evaluation reflects the quality of the system. For the subject level evaluation classification accuracy is reported, and for the final comparison of the models performance, F1 scores are reported. These metrics are reported for each class as well as for the whole test dataset.

#### 4.1.1. GANs Comparison

Table 4 shows the utterance and patient level accuracy for the classification model which used data generated by different GAN models in addition to the real data. The baseline is the model trained only on real data. The Wasserstein GAN trained for 1000 epochs demonstrates the best results on the patient level, showing 4% accuracy improvement over baseline. The Wasserstein GAN model trained for 500 epochs and simple GAN models trained for 5000 and 20000 epochs perform worse than baseline. The Wasserstein GAN trained for 2000 and the simple GAN trained for 20000 epochs demonstrates 1% improvement over baseline. The utterance and subject level accuracy for the classification model trained using only the data generated by different GAN models is shown in Table 5. The baseline is

Model	Utt. level	Subj. level
Baseline	0.45	0.44
Random Classifier	0.25	0.25
GAN 5000 epochs	0.24	0.24
GAN 20000 epochs	0.33	0.29
W-GAN 500 epochs	0.30	0.31
W-GAN 1000 epochs	0.35	0.31
W-GAN 2000 epochs	0.35	0.25

Table 5: Classification accuracy for GAN models trained for different amount of epochs (trained only on generated data)

the classification model trained only on real training data from AphasiaBank. The results for the random classification are also reported. The results show that Wasserstein GAN trained for 1000 epochs and for 500 epochs demonstrate the best improvement over random classifier, however they do not beat the baseline. The simple GAN trained for 20000 epochs also demonstrated small improvement over random classifier. Simple GAN trained for 5000 epochs and Wasserstein GAN trained for 2000 epochs show performance similar to random classifier. As Wasserstein GAN trained for 1000 epochs demonstrates the best performance in the both experiments, it is used in all the following experiments to generate synthetic data.

#### 4.1.2. Max Class Experiments

	Real	Real + Generated
MC-1	0.35	0.41
MC-2	0.38	0.48
MC-3	0.42	0.49
MC-4	0.43	<b>0.53</b>
MC-5	0.45	0.49
MC-6	0.46	0.47
MC-7	0.42	0.44

Table 6: F1 for the models trained on the real and combined data using the unsupervised Max Class models for subject level classification

Table 6 represents the results of the different Max Class models with the utterance classifier trained on real data and on combination of real and generated data. The table presents the F1 score for each of the Max Class models. The results show that the best performing model is the model trained on the combination of real and generated data on the utterance level and divides the number of the predicted non-aphasic utterance by 4. Also, the table shows that reducing the impact of the non-aphasic class helps to improve the classification.

Table 7 shows the results of the models using Max Class on the subject level for each aphasia class. The results show that none of the tested models managed to classify conduction aphasia and the models do not ever predict the conduction aphasia class. Reducing the impact of the non-aphasia class improved classification of control group and Anomic aphasia group, and did not influence classification

Model	Anom.	Broc.	Cond.	Contr.
MC-1 Real	0.41	0.34	0.00	0.64
MC-1 Comb.	0.38	0.61	0.00	0.65
MC-2 Real	0.41	0.33	0.00	0.76
MC-2 Comb.	0.46	0.60	0.00	0.87
MC-3 Real	0.44	0.33	0.00	0.88
MC-3 Comb.	0.46	0.60	0.00	0.91
MC-4 Real	0.45	0.33	0.00	0.91
MC-4 Comb	0.54	0.60	0.00	0.98
MC-5 Real	0.51	0.33	0.00	0.97
MC-5 Comb	0.51	0.60	0.00	0.87
MC-6 Real	0.51	0.33	0.00	0.98
MC-6 Comb	0.49	0.60	0.00	0.78
MC-7 Real	0.49	0.33	0.00	0.87
MC-7 Comb	0.48	0.60	0.00	0.71

Table 7: Individual F1 per class for the models trained on real and combined data using Max Class models for the subject level classification

of Broca’s and Conduction aphasia. Adding generated data improves classification of Broca’s aphasia. For Anomic aphasia, generating data improved F1 score.

#### 4.1.3. Supervised Methods Experiments

	Real	Real + Generated
Max Class	0.34	0.32
Naive Bayes	0.42	0.46
Multinomial NB	0.52	0.58
Decision Trees	0.55	0.52
Random Forest	<b>0.61</b>	0.53
KNN	0.56	0.56
SVM	<b>0.61</b>	0.58

Table 8: F1 for the models trained on the real and combined data using the supervised models for subject level classification

Table 8 shows that, although adding the generated data to the training set for utterance level classification and synthetic data generation helps when the Naive Bayes classification is used for the subject level classification improving the results of classification from F1=0.42 to F1=0.46 and F1=52 to F1=58 for Gaussian Naive Bayes and Multinomial Naive Bayes classifiers respectively, the data generation did not improve the results for other subject level classification methods. Out of all the methods used, the SVM and RF classification on the subject level without data generation showed the best results (F1= 0.61). For this methods, generating the additional data did not help to improve the classification. Also, unlike the previous experiments where both generator and utterance level classifier were trained on the whole dataset, the results for the Max Class utterance level classification did not improve when the generated data was added to the training set. For the Max Class, DT, RF and SVM subject level classification models adding the generated data made the results worse. And for the KNN classifier the F1 score stayed the same

when the generated data was included in the training set.

Model	Anom.	Broc.	Cond.	Contr.
MaxClass Real	0.42	0.19	0.00	0.75
MaxClass Comb.	0.38	0.29	0.00	0.59
NB Real	0.35	0.63	0.06	0.64
NB Comb	0.31	0.62	0.23	0.67
Mult. NB Real	0.46	0.61	0.11	0.91
Mult. NB Comb.	0.42	0.63	0.39	0.88
DT Real	0.46	0.67	0.16	0.91
DT Comb.	0.39	0.72	0.12	0.85
RF Real	0.47	0.77	0.31	0.89
RF Comb.	0.43	0.73	0.12	0.87
KNN Real	0.46	0.70	0.18	0.90
KNN Comb	0.49	0.77	0.12	0.86
SVM Real	0.53	0.75	0.18	0.97
SVM Comb.	0.47	0.74	0.18	0.94

Table 9: Individual F1 per class for the models trained on real and combined data using the supervised models for the subject level classification

Table 9 shows the results for the models trained on the real data and combination of the real and generated data for each aphasia class. It shows that unlike the Max Class methods, the supervised methods manage to sometimes predict the conduction aphasia class. However, the F1 score for this class still performed the worst out of all the classes.

## 5. Discussion

The Wasserstein GAN trained for 1000 epochs produced the best results. Wasserstein GAN trained for 500 epochs produced worse results because it probably did not converge, meaning that both discriminator and generator were not good enough to produce data resembling real data. Wasserstein GAN trained for 2000 also performed worse than the one trained for 1000 epochs. It likely means that the mode collapse problem occurred, meaning that the generator learned to produce output datapoints which were not diverse, but managed to trick the discriminator. The simple GAN trained for 5000 epochs performed the worst out of all the trained models. This model did not manage to converge, as empirically, it takes longer for the original GAN to converge due to possible oscillations in optimization, whereas WGAN has more stable training, leading to faster optimization. The simple GAN model trained for 20000 epochs performed better than the one trained for 5000 epochs. These results match our intuitions that Wasserstein GANs converge faster than simple GANs.

Supervised machine learning methods for the subject level classification outperformed unsupervised methods. Although when using the models with unsupervised subject classification, augmenting the training set with the generated samples improved the classification, the highest result for the Max Class model with the reduced impact of the non-aphasia class (F1 = 0.53) was outperformed by Multinomial NB, DT, RF, KNN, and SVM classification methods trained only on real data. The RF and SVM showed the best result. Adding the generated data to the training set improved the performance of the model only for Multino-

mial NB (from  $F1=0.52$  to  $F1=0.58$ ) and NB (from  $F1=0.42$  to  $0.46$ ) models. For the other models, including the Max Class model, the performance stayed the same (KNN) or dropped (Max Class, DT, RF, and SVM). Using the generated data did not help to beat the best performing model trained only on the real data. The reason for this may be that when using the supervised machine learning techniques on the subject level, the training data has to be split in two parts which leads to reducing the training set for generative model. It is possible, that with the reduced amount of training data the model did not manage to learn to generate samples diverse enough for helping the classification. The fact that data generation improved the performance of the simple Max Class model when trained on the whole training dataset, and failed to improve the performance of the same model when trained on the reduced dataset supports that explanation.

All tested Max Class models failed to classify Conduction aphasia and for the ML classifiers which managed to predict the Conduction aphasia class, the F1 score for the Conduction aphasia is lower than for the other classes. In AphasiaBank, Conduction aphasia has the least amount of patients. Possibly, the data was not diverse enough to classify this type of aphasia and generate good artificial data. Conduction aphasia almost always classified as Anomic aphasia. Anomic and Conduction aphasia are fairly similar in writing: both are characterised by fluent speech. In addition, the WAB aphasia severity scores for Conduction and Anomic aphasia patients are quite close, which means that these types of aphasia have similar level of severity. While patients with Anomic aphasia often use neologisms and frustration markers, patients with conduction aphasia often produce words incorrectly. In both these cases, the produced words will be treated as OOV words by the classifier and will not be accounted for.

Intuitively better classification on the utterance level should lead to the better classification on the subject level. However, this is not the case for the current experiments. The reason for this is that the test set we are evaluating the utterance level classification on is noisy, because of the aphasia patients producing non-aphasia utterances. Therefore the classification accuracy on the utterance level does not really reflect the real quality of the classification. So, there are cases when although the classification on the utterance level improves the classification on the subject level drops and other way round.

## 6. Related Work

Most of the works focused in the aphasia or mild cognitive disease classification tend to treat this problem as a binary classification problem. A lot of studies focus on the impaired and non-impaired speech classification (Järvelin and Juhola, 2011; Themistocleous et al., 2018; Little et al., 2009; Meilán et al., 2014). The others try to distinguish one type of language impairment from another, still treating the problem as binary classification (Fraser et al., 2014; Yourganov et al., 2015). Therefore, the results reported in these works cannot be directly compared to our results in the current setting. To the best of our knowledge, the classification of multiple aphasia types has not been attempted

by researchers.

However, approaches, similar to the one taken in this work, were tested in different domains and these results can be indirectly compared to ours. For example, Guan et al. (2018) used cGANs to augment training data for automatic electronic medical records (EMR) classification into diagnosis types. The task in their work is similar to ours, because they also compare the the models trained only on real data with the models trained on the combination of real and generated data. The dataset used contained 2216 EMR texts which were assigned one of the two diagnosis: pneumonia and lung cancer. For data generation, the authors use a model called Medical Text GAN (mtGAN) which generated text samples using reinforcement learning to solve the text non-differentiability problem. Guan et al. (2018) report that after adding the generated data to the real training set the classification accuracy improved from 0.7500 to 0.7635 (0.0135 improvement).

Although from the high level perspective our approach is similar to the one used by Guan et al. (2018), it differs in details. First their data contains texts written by doctors about patients, while we focus on the speech produced by the patients. Second, the different strategies are used due to the structural differences of the data; while we use two level classification, Guan et al. (2018) classify the EMRs directly. Finally, our approach to data generation is different, as we generated the data on the hidden representation level, while Guan et al. (2018) generated the textual samples. In the case of the unsupervised subject classification, our results demonstrate bigger improvement when using generated data in combination with real data. The best system using Max Class system and only real data demonstrate the accuracy of 0.46, while adding the generated data brings the accuracy to 0.53 (0.07 improvement). The bigger improvement in the aphasia classification case could be caused by the difference in the approaches as well as by the difference in the datasets.

## 7. Conclusion

The method of using the same text representations for both generation and classification tasks was proposed. By encoding the text into vectors from the beginning and then generating and classifying vector representations, we avoid the problem of text being discrete when using GANs. Also, this approach requires only encoding the text, but no decoding is needed.

The results show that using hidden feature generation with GANs for improving text classification is useful in certain cases, and generating additional synthetic data and combining it with the real data for training improves the classification results. However, it has certain limitations, namely, the generation model still needs sufficient amount of data to be able to produce useful output.

## 8. Acknowledgements

This work was partially funded by the European Union's Horizon 2020 grant agreement No. 777107 (Precise4Q) and by the BMBF project DeepLee (01IW17001).

## 9. Bibliographical References

- Allen, L., Mehta, S., Andrew McClure, J., and Teasell, R. (2012). Therapeutic interventions for aphasia initiated more than six months post stroke: a review of the evidence. *Topics in stroke rehabilitation*, 19(6):523–535.
- Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein gan. arxiv e-prints, page. *arXiv preprint arXiv:1701.07875*.
- Basso, A., Taborelli, A., and Vignolo, L. (1978). Dissociated disorders of speaking and writing in aphasia. *Journal of Neurology, Neurosurgery & Psychiatry*, 41(6):556–563.
- Bhagal, S. K., Teasell, R., Speechley, M., and Albert, M. (2003). Intensity of aphasia therapy, impact on recovery. *Stroke—a Journal of Cerebral Circulation*, 34(4):987–991.
- Bond, M. and Brooks, D. (1976). Understanding the process of recovery as a basis for the investigation of rehabilitation for the brain injured. *Scandinavian Journal of Rehabilitation Medicine*.
- Broca, P. (1861). Remarks on the seat of the faculty of articulated language, following an observation of aphemia (loss of speech). *Bulletin de la Société Anatomique*, 6:330–57.
- Caramazza, A. (1984). The logic of neuropsychological research and the problem of patient classification in aphasia. *Brain and language*, 21(1):9–20.
- Chen, J., Chun, D., Patel, M., Chiang, E., and James, J. (2019). The validity of synthetic clinical data: a validation study of a leading synthetic data generator (synthea) using clinical quality measures. *BMC medical informatics and decision making*, 19(1):44.
- Damasio, A. R. (1992). Aphasia. *New England Journal of Medicine*, 326(8):531–539.
- Demeurisse, G., Demol, O., Derouck, M., De Beuckelaer, R., Coekaerts, M., and Capon, A. (1980). Quantitative study of the rate of recovery from aphasia due to ischemic stroke. *Stroke*, 11(5):455–458.
- Dziugaite, G. K., Roy, D. M., and Ghahramani, Z. (2015). Training generative neural networks via maximum mean discrepancy optimization. *arXiv preprint arXiv:1505.03906*.
- Fraser, K. C., Meltzer, J. A., Graham, N. L., Leonard, C., Hirst, G., Black, S. E., and Rochon, E. (2014). Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. *cortex*, 55:43–60.
- 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.
- Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., and Greenspan, H. (2018). Synthetic data augmentation using gan for improved liver lesion classification. In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 289–293. IEEE.
- Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. (2016). Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.
- Goodfellow, I. (2016). Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*.
- Guan, J., Li, R., Yu, S., and Zhang, X. (2018). Generation of synthetic electronic medical record text. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 374–380. IEEE.
- Hanson, W. R., Metter, E. J., and Riege, W. H. (1989). The course of chronic aphasia. *Aphasiology*, 3(1):19–29.
- Huszár, F. (2015). How (not) to train your generative model: Scheduled sampling, likelihood, adversary? *arXiv preprint arXiv:1511.05101*.
- Järvelin, A. and Juhola, M. (2011). Comparison of machine learning methods for classifying aphasic and non-aphasic speakers. *Computer methods and programs in biomedicine*, 104(3):349–357.
- Jefferies, E. and Lambon Ralph, M. A. (2006). Semantic impairment in stroke aphasia versus semantic dementia: a case-series comparison. *Brain*, 129(8):2132–2147.
- Jennett, B. and Bond, M. (1975). Assessment of outcome after severe brain damage: a practical scale. *The Lancet*, 305(7905):480–484.
- Kertesz, A. (2007). *WAB-R: Western aphasia battery-revised*. PsychCorp.
- Kinsella, G. and Ford, B. (1980). Acute recovery from patterns in stroke patients: neuropsychological factors. *The Medical Journal of Australia*, 2(12):663–666.
- Kolk, H. (1998). Disorders of syntax in aphasia: Linguistic-descriptive and processing approaches. In *Handbook of neurolinguistics*, pages 249–260. Elsevier.
- LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551.
- Leung, J. H., Purdy, S. C., Tippett, L. J., and Leão, S. H. (2017). Affective speech prosody perception and production in stroke patients with left-hemispheric damage and healthy controls. *Brain and language*, 166:19–28.
- Li, Y., Swersky, K., and Zemel, R. (2015). Generative moment matching networks. In *International Conference on Machine Learning*, pages 1718–1727.
- Little, M. A., McSharry, P. E., Hunter, E. J., Spielman, J., and Ramig, L. O. (2009). Suitability of dysphonia measurements for telemonitoring of parkinson’s disease. *IEEE transactions on bio-medical engineering*, 56(4):1015.
- MacWhinney, B., Fromm, D., Forbes, M., and Holland, A. (2011). Aphasiabank: Methods for studying discourse. *Aphasiology*, 25(11):1286–1307.
- Maqsd, U. (2015). Synthetic text generation for sentiment analysis. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 156–161.

- Marshall, J., Atkinson, J., Smulovitch, E., Thacker, A., and Woll, B. (2004). Aphasia in a user of british sign language: Dissociation between sign and gesture. *Cognitive neuropsychology*, 21(5):537–554.
- Meilán, J. J. G., Martínez-Sánchez, F., Carro, J., López, D. E., Millian-Morell, L., and Arana, J. M. (2014). Speech in alzheimer’s disease: Can temporal and acoustic parameters discriminate dementia? *Dementia and Geriatric Cognitive Disorders*, 37(5-6):327–334.
- Mesulam, M.-M., Thompson, C. K., Weintraub, S., and Rogalski, E. J. (2015). The wernicke conundrum and the anatomy of language comprehension in primary progressive aphasia. *Brain*, 138(8):2423–2437.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Mirza, M. and Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.
- Nguyen, A., Clune, J., Bengio, Y., Dosovitskiy, A., and Yosinski, J. (2017). Plug & play generative networks: Conditional iterative generation of images in latent space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4467–4477.
- Nowozin, S., Cseke, B., and Tomioka, R. (2016). f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in neural information processing systems*, pages 271–279.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.
- Purdy, M., Coppens, P., Madden, E. B., Mozeiko, J., Patterson, J., Wallace, S. E., and Freed, D. (2019). Reading comprehension treatment in aphasia: A systematic review. *Aphasiology*, 33(6):629–651.
- Qin, Y., Lee, T., Feng, S., and Kong, A. P. H. (2018). Automatic speech assessment for people with aphasia using tdnn-blstm with multi-task learning. *Proc. Interspeech 2018*, pages 3418–3422.
- Risser, A. H. and Spreen, O. (1985). The western aphasia battery. *Journal of clinical and experimental neuropsychology*, 7(4):463–470.
- Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242.
- Skilbeck, C. E., Wade, D. T., Hewer, R. L., and Wood, V. A. (1983). Recovery after stroke. *Journal of Neurology, Neurosurgery & Psychiatry*, 46(1):5–8.
- Swindell, C. S., Holland, A. L., and Fromm, D. (1984). Classification of aphasia: Wab type versus clinical impression. In *Clinical Aphasiology: Proceedings of the Conference 1984*, pages 48–54. BRK Publishers.
- Themistocleous, C., Eckerström, M., and Kokkinakis, D. (2018). Identification of mild cognitive impairment from speech in swedish using deep sequential neural networks. *Frontiers in neurology*, 9.
- Ulatowska, H. K., North, A. J., and Macaluso-Haynes, S. (1981). Production of narrative and procedural discourse in aphasia. *Brain and language*, 13(2):345–371.
- Yourganov, G., Smith, K. G., Fridriksson, J., and Rorden, C. (2015). Predicting aphasia type from brain damage measured with structural mri. *Cortex*, 73:203–215.
- Zhao, J., Mathieu, M., and LeCun, Y. (2016). Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*.

# Automated Analysis of Discourse Coherence in Schizophrenia: Approximation of Manual Measures

**Galina Ryazanskaya, Mariya Khudyakova**

National Research University Higher School of Economics  
Moscow, 21/4 Staraya Basmannaya str., Center for Language and Brain  
galka1999@gmail.com, mariya.kh@gmail.com

## Abstract

Disorganized, or incoherent, speech is one of the important criteria for diagnosing schizophrenia. However, there is still a lack of a rather quick objective method of measuring speech coherence. Automated discourse analysis is a possible solution to this problem. We analyzed discourse coherence in a set of spoken narratives by people with schizophrenia and neurotypical speakers of Russian. All narratives were manually rated for violations of completeness, local, global and dimensional coherence. A number of automated vector semantics methods were used for approximation of the manual rating scores. The metrics used proved to be a good approximation for manual scoring, and a combination of them was efficient for classification narratives in schizophrenia and neurotypical groups.

**Keywords:** schizophrenia, discourse coherence, vector semantics

## 1. Introduction

### 1.1 Schizophrenia and discourse coherence

Schizophrenia is a severe mental illness characterized by the presence of at least some of the following symptoms: cognitive and perceptual impairments such as auditory and visual hallucinations, delirium, disordered thought and speech, and inadequate affect (Ditman & Kuperberg, 2010; Harrow & Jobe, 2010). Disorganized speech in schizophrenia is believed to be reflective of disorganization in the thought process, such as the loosening of associations (Bleuler, 1911/1950; Kraepelin, 1921). It is included as one of the main symptoms of schizophrenia in the two most widely used psychiatric diagnostic manuals, ICD-10 (WHO, 1992) and DSM-5 (Tandon et al., 2013).

However, “disorganized” or “incoherent” speech is problematic as a diagnostic criterion given the lack of a clear definition or reliable assessment methods. In this paper, we discuss different coherence measures that can be applied to measure the level of speech disorganization and propose an automated method for approximating expert evaluation and manual metrics. Below we will discuss approaches to defining and measuring coherence and speech organization in psychiatry and clinical linguistics, give an overview of automated methods of language analysis used for assessment of speech in populations with psychiatric disorders and provide details about the most relevant studies on automated analysis of speech disorganization.

Psychiatric assessment of speech coherence in clinical practice is highly subjective, as neither of the diagnostic manuals defines speech disorganization. The definitions in research are plentiful, and no single one is used consistently in practice (see Ditman & Kuperberg, 2010). Besides, “disorganized speech” as a symptom lacks linguistic insight (Cohen et al., 2017), as it fails to reflect the fact that language has multiple interdependent levels of organization (phonetics, morphology, syntax, discourse, pragmatics, and interactional markers). There are standard linguistic measures of discourse organization or coherence; these, however, suffer from similar pitfalls (subjectivity) and, on top of that, require time-consuming rating by trained linguists. Automated methods, although initially reliant on a ‘gold standard’, guarantee high reproducibility.

### 1.2 Automated analysis of speech in schizophrenia

Automated discourse analysis is a well-developed field of computational linguistics that uses computational methods to extract discourse-level features from a text. The tasks range from shallow, low-level parsing, such as coreference (or anaphora) resolution, to very high-level ones that seek to approximate the overall structure of discourse.

In recent years, there has been a growing interest in applying these methods to the discourse by patients with various mental disorders to add an objective measure to psychiatric speech-based diagnosis (see Qiwei He, 2013; Cohen & Elvevaag, 2014; Abbe et al., 2016 for review).

Distributional or vector semantics is a family of mechanisms that allow for vector representation of words, that are called word embeddings. The algorithms are such that words that occur in the same context are close in the resulting multidimensional space. The measure of the proximity of the words is called cosine similarity. There are two types of embeddings. They can be context-independent, meaning every word always has the same vector. Such models include Latent Semantic Analysis or LSA (Foltz et al., 1998), word2vec or w2v (Mikolov et al., 2013), and GloVe (Pennington et al., 2014). The other more modern type of embeddings is context-dependent embeddings, which means that the vector of a given word depends on the surrounding words in the sentence. This method is more complicated, but it represents homonyms and polysemic words with different vectors, depending on the meaning in the context. These include ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018).

Elvevaag and colleagues (2007) were the first to apply the Latent Semantic approach to assess discourse organization in schizophrenia. A set of responses in the structured interview - descriptions of the Cinderella story and of the process of doing laundry - were rated by two psychiatrists on a Likert scale from 1 (‘very coherent’) to 7 (‘bizarre, completely incoherent’). An average vector of each sentence in the stories and the centroid were calculated so that every story was represented by one vector. Next, cosine distances between each participant’s response vector and the centroid of all the responses by all participants were computed. The researchers found that the correlation coefficients between the coherence estimates

yielded by the vector approximation and by the experts amounted to 0.7 for the Cinderella story and 0.5 for the laundry story on average across three coherence subdimensions used.

The LSA model used by Elvevaag and colleagues (2007) was trained on a corpus of written texts, which may not be the best fit for spoken-language applications. To overcome these limitations, later studies used Internet corpora as a better approximation of spoken language (Iter et al, 2018; Just et al, 2019; Panicheva & Litvinova, 2019), and it was suggested that fillers and repetitions should be omitted (Iter et al, 2018). Despite the fact that self-repairs and filled pauses can be predictive of the diagnosis (Howes et al 2017), they cannot be treated by the models the same way the words are, and thus should be excluded from this kind of automated analysis.

Despite progress in the area, the questions about the best vector semantics measures to assess coherence remain open. For example, while Iter and colleagues (2018) found that four out of 20 models tested were able to distinguish between the clinical groups. Just and colleagues (2019) did not replicate the results on a larger group of speech samples, as they found only one model to have predictive power. This difference might be partially explained by the smaller training corpus available for German, as well as by the fact that German morphology is richer than English. German also allows the use of complex compounds that cannot be recognized by the model, which brings down the model performance.

One of the other important questions about the automated assessment of coherence is whether the metrics yield the same results when applied to different languages. For example, when an adapted version of the semantic coherence method from Elvevaag et al. (2007) was applied to written texts in Russian, it was able to successfully classify texts by schizophrenia group versus the control group with relatively high accuracy (0.72-0.88). However, some of the results were contradictory to the original results obtained for the English language. For example, the minimum semantic coherence was higher in schizophrenic texts than in control ones, while Elvevaag and colleagues observed higher coherence scores in the neurotypical population (Panicheva & Litvinova, 2019).

### 1.3 Present study

In this paper, we try to bridge the gap between computational linguistics and psychiatric diagnosis, by approximating “manual” linguistic measures with vector analogs on a clinical sample, as well as a control group. Previous work revealed that lexical variability might strongly influence the results as vector semantic methods are very sensitive to lexical choice and out-of-vocabulary issues. To ensure a clear narrative structure, as well as restricted lexical variability, we used film retellings as the elicitation technique. Thus, we will take into account the insight from the previous research, i.e. excluding filler words and repetitions, as well as using w2v trained on an internet corpus.

## 2. Material

### 2.1 Participants

9 (3 female) outpatients diagnosed with schizophrenia (3 females; mean age – 35; age range – 22-53) were recruited for participation in the study. The only psychiatric data available for the patient group was the diagnosis. Data for the neurotypical control group (5 females; mean age – 58; age range – 25-78) were taken from the Russian ClipS corpus (Khudyakova et al., 2016). All participants had normal or corrected to normal vision and hearing. All have given their informed consent.

### 2.2 Procedure

Participants asked to watch a short Pear film (Chafe, 1980) and retell the plot of the film in such a way that someone who has not seen the film could understand what had happened in it. The Pear film is a 6-minute-long speechless film created specifically for elicitation of narratives across different languages and cultures. The retellings were transcribed, annotated and segmented into elementary discourse units (EDUs) in ELAN (Wittenburg et al., 2006). The detailed description of the procedure and annotation scheme can be found in Khudyakova et al. (2016).

## 3. Coherence annotation

There are several established measures of discourse organization. The ones selected for the study are discussed in greater detail below.

### 3.1 Manual annotation

The transcripts of retellings were manually scored for violations of completeness (Christiansen, 1995), local and global coherence (Glosser & Deser, 1990 + Coelho & Flewellyn, 2003), and dimensional measure of discourse coherence.

#### 3.1.1 Violations of Completeness

To evaluate the informativeness aspect of the retellings, we used the number of violations of completeness (Christiansen, 1995). To implement the measure, one identifies the important propositions of the original story and then for a given retelling computes the number of missing important propositions. The list of propositions was based on the plot of the elicitation stimulus, the Pear film (Chafe, 1980). The list of propositions can be found at the project's [GitHub repository](#)<sup>1</sup>. Violations of completeness were rated by 5 diagnosis-blind annotators with at least 3 years of formal linguistic training rated. The annotators were asked to mark each proposition as either absent or present in a retelling. We have found low (Cohen's kappa < 0.5) inter-rater agreement.

#### 3.1.2 Global and Local Coherence

Global coherence can be defined as a relationship of every utterance to the overall topic of the text at hand, while local coherence reflects the similarity in content and logical connectedness of two adjacent utterances. Measures of global and local coherence were developed for the analysis of discourse in aphasia (Glosser & Deser, 1990) and have not been applied to speech in schizophrenia (Ditman & Kuperberg, 2010). The transcripts were annotated by the

<sup>1</sup> [https://github.com/flying-bear/modeling\\_schizo](https://github.com/flying-bear/modeling_schizo)

author. Each EDU received a score for global and local coherence on the 5-point scale according to the instruction provided in (Wright, Capilouto, & Koutsoftas, 2013).

### 3.1.3 Dimensional Local coherence

Using Likert-type scales for evaluating coherence, while being very common in clinical linguistics, has crucial limitations. First, higher numbers of alternatives might be less reliable, while binary options tend to be very easy to reproduce, obtaining the same result (Matell & Jacoby, 1971). Second, statistical analysis of Likert-type scales is somewhat controversial, because they are ordinal categorical variables but are often treated as discrete or even continuous (for example, see Harpe, 2015). Also, in many cases, the scores are distributed neither normally nor uniformly, which significantly restricts the range of possible statistical tools. To combat some of these limitations, we introduced a dimensional method of assessing discourse coherence.

Dimensional local coherence is an alternative way of assessing coherence that avoids using a Likert-type scale. As there are many ways of maintaining connectedness of adjacent clauses, each of the methods can be used as an indication of local coherence. We used a binary (present-absent) score for each of the 3 dimensions of local coherence. The transcripts were annotated by the author.

The dimensions identified were as follows:

1. *Topic* – Utterance containing a repeated nominal phrase or an unambiguous anaphoric reference to the previous utterance, was marked as having the same “topic”.
2. *Time* – Utterance that was either following the previous one in the time frame of the plot or was describing a simultaneous event, as the previous utterance, was marked as having a locally coherent time.
3. *Discourse level* – Three possible discourse levels were identified. Each utterance could contain each of the levels. If adjacent utterances shared at least one discourse level, they were marked as locally coherent with respect to the discourse level. The sub-dimensions (SD) identified were

- 3.1. *Story* – Events and scenes of the film or things that the participant believes to be present in the film.
- 3.2. *Comment* – The participants’ attitudes towards the events and characters of the film, as well as possible interpretations of the events.
- 3.3. *Meta-comment* – Things unrelated to the plot: comments on the film, one’s speech or experimental setting, details of one’s biography, unfinished utterances of unclear discourse level.

### 3.2 Automated Annotation

All Python and R code for this project can be found at a [GitHub repository](#)<sup>2</sup>.

The semantic vector space model was trained on a large (140 million words) Russian Internet corpora [RuWac](#)<sup>3</sup> (Khokhlova, 2004). RuWac is a corpus of written internet speech, and thus it can be argued to be closer to spoken corpus, than a regular written corpus. There exists no corpus of spoken Russian large enough for training a word embedding model.

As Russian is a highly inflected language, all the words were lemmatized. The corpus was plotted into sentences and all the sentences with Latin characters were excluded. The previously collected texts underwent similar operations, except utterances were used instead of sentences and lemmatization was done manually. The internet corpus was used as the closest approximation of a large enough corpus of spoken language. The texts in the corpus and the sample texts were combined and fed into the gensim w2v as training data with the following parameters: 300 dimensions, 1 epoch (due to limited computational power), skip-gram algorithm, minimal word count at 10 (all other parameters unspecified, set to [standard](#)<sup>4</sup>). The resulting vector space was TF-IDF weighted to account for low semantic informativeness of the most frequent words. The application of TF-IDF weighting for vector embeddings is discussed in Lintean et al. (2010).

The Tables 1 and 2 below contain formulae, abbreviations, and explanations of all the automated metrics used in this study.

<sup>2</sup> [https://github.com/flying-bear/modeling\\_schizo](https://github.com/flying-bear/modeling_schizo)

<sup>3</sup> <https://www.sketchengine.eu/russian-web-corpus/>

<sup>4</sup> <https://radimrehurek.com/gensim/models/word2vec.html>

name	notation	formula	explanation
cosine similarity	$\cos(v_1, v_2)$	$\cos(v_1, v_2) = \frac{v_1 \cdot v_2}{ v_1  \cdot  v_2 }$ , where $v_1$ and $v_2$ are some word vectors.	The proximity of two words according to the model (i.e. the angle between their vectors).
inverse document frequency	$idf(w)$	$idf(w) = \log\left(\frac{ D }{ \{utt_i \in D \mid w \in utt_i\} }\right)$ , where $ D $ is the size of the collection in utterances, and $ \{utt_i \in D \mid w \in utt_i\} $ is the number of utterances in the collection, containing the word $w$ . In our particular case, $ D $ was set to sample size (627 utterances) + first 100000 utterances from RuWac.	A measure of word “meaningfulness”: log of (the number of utterances in the collection divided by the number of utterances in the corpus containing this word).
utterance vector	$v_{utt}$	$v_{utt} = \frac{\sum_{i=1}^{ utt } idf(w_i) v_i}{ utt  + \sum_{i=1}^{ utt } idf(w_i)}$ , where $ utt $ is the length of the utterance in words, $w_i$ is the $i^{\text{th}}$ word of the utterance, and $v_i$ is the vector of this word Words absent in the model were skipped.	An IDF-weighted average of word vectors.
cumulative vector	$v_{cum}(n)$	$v_{cum}(n) = \frac{\sum_{i=1}^n v_{utt,i}}{n}$ , where $v_{utt,i}$ is the vector of the $i^{\text{th}}$ utterance.	An average of all the story vectors up to the $n^{\text{th}}$ .
story vector	$v_{story}$	$v_{story} = v_{cum}( story ) = \frac{\sum_{i=1}^{ story } v_{utt,i}}{ story }$ , where $ story $ is the length of the story in utterances, and $v_{utt,i}$ is the vector of the $i^{\text{th}}$ utterance.	The average of utterance vectors.
propositions vector	$v_{prop}$	$v_{prop} = v_{story}   story = \{plot\ propositions\}$ , which is $v_{prop} = \frac{\sum_{i=1}^{ prop } v_{utt,i}}{ prop }$ , where $ prop $ , the length of the set of propositions, is 7, and $v_{utt,i}$ is the vector of the $i^{\text{th}}$ proposition.	The average of the utterance vectors of the plot propositions list.
sample stories vector	$v_{all}$	$v_{all} = \frac{\sum_{i=1}^{ sample } v_{story,i}}{ sample }$ , where $ sample $ is the sample size in stories, and $v_{story,i}$ is the vector of the $i^{\text{th}}$ story. In our particular case, $ sample $ is 19.	The average of all the story vectors in the sample
control stories vector	$v_{control}$	$v_{control} = \frac{\sum_{i=1}^{ control } v_{story,i}}{ control }$ , where $ control $ is the size of the control group in stories, and $v_{story,i}$ is the vector of the $i^{\text{th}}$ story In our particular case, $ control $ is 10.	The average of all the story vectors in the control group.

Table 1: The names and the formulae of the vectorization methods used.

name	metric approximated	formula	explanation
auto_compl_viol_inverse	completeness violations (inverse)	$\cos(v_{story}, v_{prop})$	The cosine distance between the vector of the topics list and the story vector.
auto_glob_coh	global coherence	$\cos(v_{story}, v_{all})$	The cosine distance between the vector of all the stories and the vector of the current story (adapted from Elvevaag et al., 2007).
auto_glob_coh_control	global coherence	$\cos(v_{story}, v_{control})$	The cosine distance between the vector of all the stories of the control group and the vector of the current story (adapted from Elvevaag et al., 2007).
auto_loc_coh	local coherence	$\frac{\sum_{i=1}^{ story -1} \cos(v_{utt,i}, v_{utt,i+1})}{ story -1}$ , where $ story $ is the length of the story in utterances.	The cosine distance between the vectors of two adjacent utterances. The final metric is the average across the utterances in a story.
auto_cum_semi_loc_coh	—	$\frac{\sum_{i=1}^{ story -1} \cos(v_{cum}(i), v_{utt,i+1})}{ story -1}$ , where $ story $ is the length of the story in utterances.	The cosine distance between an averaged vector of the first n clauses and the current clause. The final metric is the average across n ranging from 1 to story length – 1.

Table 2: The names and the formulae of the automated metrics used to approximate the manual measures of discourse coherence.

## 4. Preliminary Results

### 4.1 Descriptive Statistics

We ran a Pearson’s correlation test using R. No correlation with gender (p-value > 0.05), age (p-value = 0.05 for completeness violations, and p-value > 0.05 for other metrics) or education level (p-value > 0.05) was found for the classic manual measures of discourse coherence (as indicated by Pearson’s correlation coefficient).

Overall, the control group produced more lengthy retellings (ranging from 15 to 57 utterances, averaging at 38, as compared to range from 9 to 50, averaging at 27 utterances in participants with schizophrenia). The difference between these lengths does not reach significance ( $t = 1,909$ ; p-value = 0.07). However, this is still important, as some coherence measures were shown to be sensitive to the length of the

text (Just et al., 2019). The lower verbosity as compared to controls has been observed in schizophrenia by many researchers independently (Just et al., 2019; for a review of language in schizophrenia see Kuperberg, 2010).

### 4.2 Predictive Power

The biplot in Figure 1 (a Principal Component Analysis of all the metrics, except the number of utterances) shows that it is possible to meaningfully split the sample into patient-control parts with only one false-positive result. The dimensions loosely correspond to dimensional local coherence: the switch dimension (accounts for 58.7% of variance) and “discourse level integration” (21.7% of variance). “Discourse level integration” is the dimension on which comment SD is very high and meta-comment SD is very low.

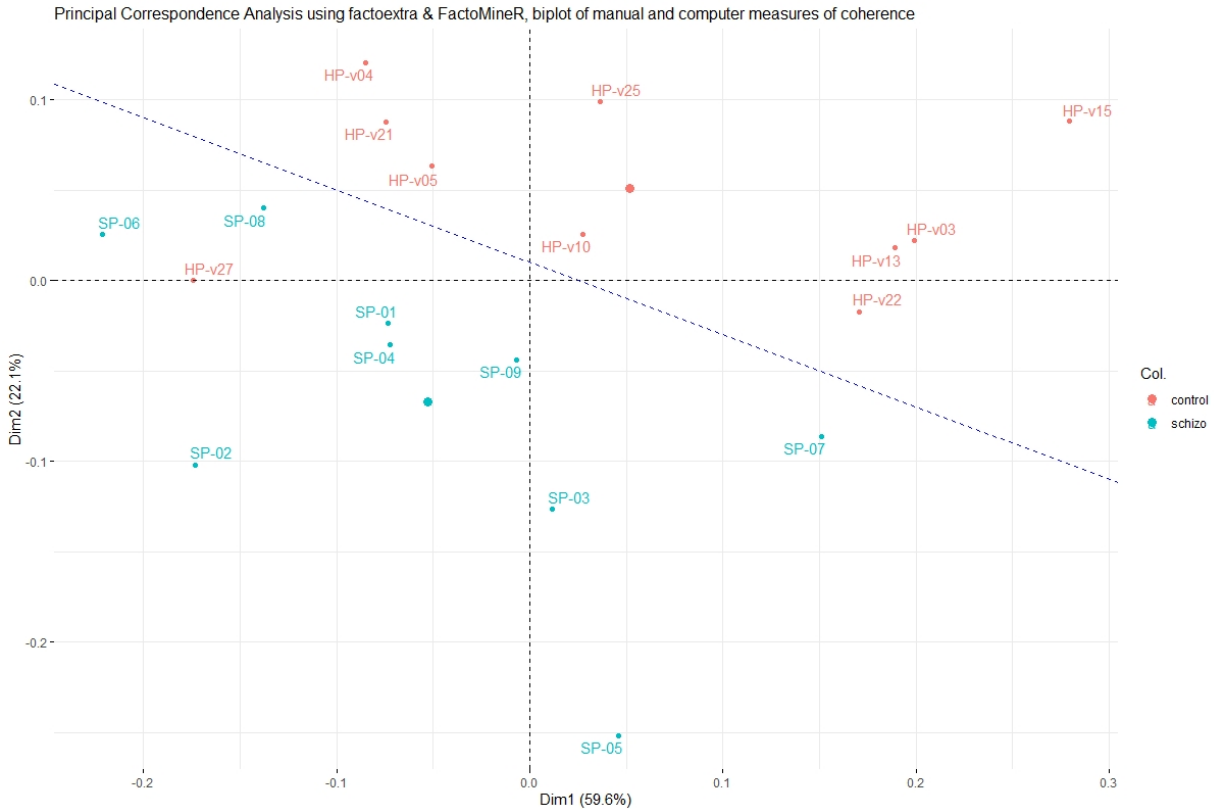


Figure 1. PCA biplot of all the metrics (excluding the number of utterances) showing the participants. The patients are shown in blue, and the controls are in red. The blue dashed line shows the optimal clustering split, which is equally affected by both dimensions.

### 4.3 Automated Approximation

The table in figure 2 shows the strength (effect size) of Pearson's correlation across metrics. Only significant (at  $p < 0.01$ ) results are shown, if the p-value is not otherwise specified it is below this threshold.

Manual metrics were largely intercorrelated.

Meta-comment SD correlated negatively with story ( $r = -0.84$ ), time ( $r = -0.81$ ), and person ( $r = -0.79$ ) SD, as well as with the dimensional local coherence as a whole ( $r = -0.78$ ). It was also negatively correlated with manual measures of local ( $r = -0.66$ ) and global ( $r = 0.63$ ) coherence.

Switch dimension correlated negatively with story SD ( $r = -0.76$ ) and dimensional local coherence as a whole ( $r = -0.78$ ). It was positively correlated with comment SD ( $r = 0.61$ ). Story SD and time dimension were all correlated positively with dimensional local coherence ( $r = 0.9$  and  $r = 0.91$  respectively), as well as manual measures of local ( $r = 0.66$  and  $r = 0.64$ ) and global coherence ( $r = 0.6$  and  $r = 0.62$ ) which were also inter-related ( $r = 0.86$ ). Time dimension was also correlated with person dimension ( $r = 0.83$ ). Person dimension in turn was also correlated with all

the manual coherence measures: local ( $r = 0.81$ ), global ( $r = 0.6$ ), and dimensional ( $r = 0.93$ ).

Completeness violations correlated with global ( $r = 0.69$ ), local ( $r = 0.69$ ). Only the approximation of global coherence on all the texts correlates with the number of utterances ( $r = 0.58$ ).

All approximation metrics positively correlated with each other with effect sizes ranging from  $r = 0.52$  ( $p < 0.05$ ) for the approximation of completeness violations and local coherence to  $r = 0.91$  for the two ways of approximating global coherence.

Finally, the approximation metrics correlated with manual metrics. Computational approximation of completeness violations was correlated positively with completeness violations ( $r = 0.66$ ), global ( $r = 0.83$ ) and local ( $r = 0.8$ ) coherence. Semi-local coherence correlated positively with local coherence ( $r = 0.67$ ) and completeness violations ( $r = 0.58$ ). Computational local coherence correlated with comment dimension ( $r = 0.63$ ), that was predictive of the diagnosis (Student  $t = 3.93$ ,  $df = 14.5$ ,  $p\text{-value} < 0.0015$ ). Approximated control global coherence correlated with local coherence ( $r = 0.62$ ).

Metrics correlation plot at significance level 0.01

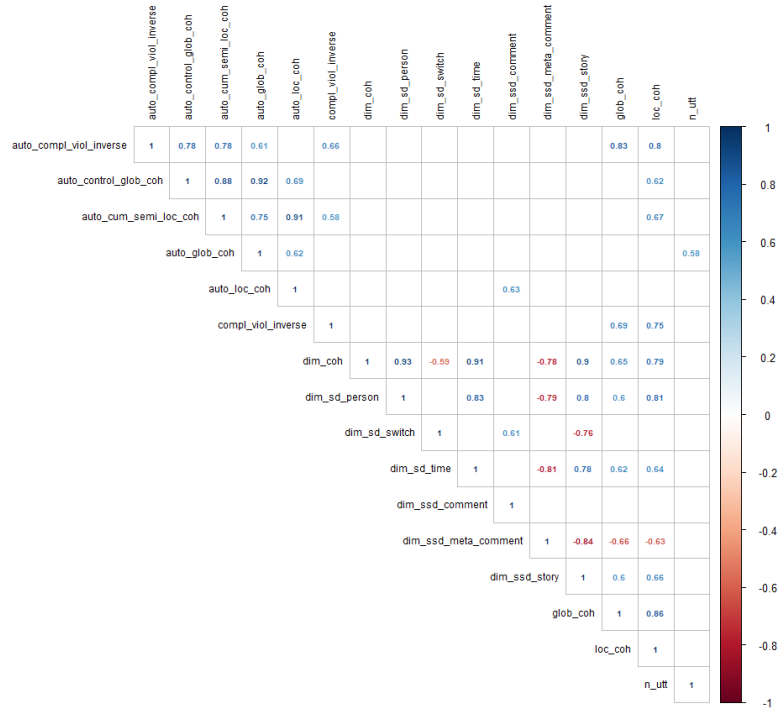


Figure 2: Pearson's correlation coefficient matrix for all the measures explored. The color scale shows the effect size (blue for positive correlations, red for negative correlations). The darker the color, the greater is the effect. All values below the 0.01 significance level are left white.

Note. Automatic measures: auto\_compl\_viol\_inverse – cosine distance between the vector of the topics list and the story vector, auto\_control\_glob\_coh – cosine distance between the vector of all the stories in the control group and the vector of the current story, auto\_cum\_semi\_loc\_coh – cosine distance between an averaged vector of the first n clauses and the current clause, auto\_glob\_coh – cosine distance between the vector of all the stories and the vector of the current story, auto\_loc\_coh – cosine distance between the vectors of two adjacent utterances. Manual measures: compl\_viol\_inverse – completeness violations (inverse), dim\_coh - dimensional local coherence, dim\_sd\_person – person dimension SD, dim\_sd\_switch – dimension switch SD, dim\_sd\_time – time dimension SD, dim\_ssd\_comment – comment dimension SD, dim\_ssd\_meta\_comment – meta-comment dimension SD, dim\_ssd\_story – story dimension SD, glob\_coh – global coherence, loc\_coh – local coherence, n\_utt – number of utterances.

## 5. Discussion

### 5.1 Predictive Power

The only manual metric that correlated with the diagnosis was the comment SD of dimensional local coherence. This might indicate that people with schizophrenia use fewer clauses that relate the speaker to the content of their speech. This, however, requires further exploration.

A very promising result is the fact that PCA is able to classify retellings based on the diagnosis with a precision of 0.9 and recall of 0.9 (F1-measure of 0.9474), i.e. one false-positive in 19 classified retellings. This might indicate that a combination of metrics (especially dimensional local coherence) is a promising direction for research.

### 5.2 Manual Annotation

The correlations within manual metrics reveal the internal structure of the dimensional local coherence. The negative correlation of meta-comment SD with manual coherence

measures is expected, as meta-comment was defined to be unrelated to the plot of the pear film. The negative correlation of switch SD with story SD and dimensional local coherence is also understandable, as switch reflects the amount of abrupt changes in discourse level. However, many shifts in the discourse level were smooth, as adjacent sentences could have one shared discourse level out of three, and the switch would be marked as absent. Thus, some more smooth discourse level changes were not reflected by the switch SD.

### 5.3 Automated Approximation

Unexpectedly, only one metric, namely, approximated global coherence correlates with the number of utterances. It is a nice result, as patients with schizophrenia are known to produce shorter texts (Just et al., 2019; for a review of language in schizophrenia see Kuperberg, 2010). However, as most metrics are not correlated with the number of utterances, there is no need for additional control of this factor.

The inter-correlatedness of the approximation metrics is explained by the fact that they are all a function of the same cosine similarity modeling. The correlation matrix shows that an approximation of manual measures of coherence with automated metrics is possible. Approximation of completeness violations and approximation of semi-local coherence are correlated with the most manual metrics and might be prominent approximations as they are independent of the number of utterances. The correlation of computational local coherence with the comment SD is an unexpected result. This might be explained by the fact that both of these are somewhat higher in controls than patients. The correlation of approximated control global coherence with manual local coherence requires further investigation.

#### 5.4 Limitations

The most significant limitation of this study is the small sample size, which might have undermined statistical robustness of the study, introducing random variation and lacking statistical power, given the true effect size might be small. The other limitation is the absence of psychiatric data beyond diagnosis: some of the patients, although suffering from schizophrenia, did not exhibit disordered speech symptoms, and thus the automated methods would not be able to detect it. In addition, though we acknowledge the

As for the algorithmic part of the study, skipping the words absent in the model (out-of-vocabulary issues) downplays the importance of neologisms that might be more apparent in disordered speech. Finally, averaging vectors across sentences ignores word order which is a very important indicator of syntactic coherence. It had to be left outside the scope of this study, which is focused on the higher-level discourse coherence.

#### 5.5 Further research

Our findings are in line with previous research on automated coherence assessment in schizophrenia: people with the diagnosis produce retellings with lower overall computational and manual coherence metrics (even if statistically insignificant). A larger sample size might help to shed light on whether automated metrics can be a predictive metric for the diagnosis.

A good technique to implement in further studies would be to use context-dependent word embeddings such as ELMo or BERT, rather than word2vec. These methods of word embeddings would alter the vectorization depending on the context and would resolve some of the out-of-vocabulary issues. Another proposal for further research is to use some sentence averaging like the one used in Iyer et al. (2018), namely, SIF, rather than a bag of words.

### 6. Conclusion

This paper is the first study using computational linguistics for psychiatric diagnosis on Russian material. The methods used in the paper allow successful approximation of linguistic measures of discourse incoherence by theoretically-driven automated metrics. A combination of the metrics can be used for relatively good diagnosis classification. The methods presented in the paper might be further developed into clinical software.

### 7. Acknowledgements

This work was funded within the framework of a subsidy by the Russian Academic Excellence Project '5-100'.

### 8. Bibliographical References

- Abbe, A., Grouin, C., Zweigenbaum, P., & Falissard, B. (2016). Text mining applications in psychiatry: A systematic literature review. *International Journal of Methods in Psychiatric Research*, 25(2), 86–100. <https://doi.org/10.1002/mpr.1481>
- Bleuler, E. (1911). Dementia praecox: oder Gruppe der Schizophrenien. F. Deuticke.
- Bois, J. W. Du, & Barbara, S. (1980). Cognitive, Cultural, and Linguistic Aspects of Narrative Production The Trace of Identity in Discourse.
- Cohen, A. S., & Elvevåg, B. (2014). Automated computerized analysis of speech in psychiatric disorders. *Current Opinion in Psychiatry*. Lippincott Williams and Wilkins. <https://doi.org/10.1097/YCO.0000000000000056>
- Cohen, A. S., Le, T. P., Fedechko, T. L., & Elvevåg, B. (2017). Can RDoC help find order in thought disorder? *Schizophrenia Bulletin*. <https://doi.org/10.1093/schbul/sbx030>
- Ditman, T., & Kuperberg, G. R. (2010). Building coherence: A framework for exploring the breakdown of links across clause boundaries in schizophrenia. *Journal of Neurolinguistics*, 23(3), 254–269. <https://doi.org/10.1016/j.jneuroling.2009.03.003>
- Elvevåg, B., Foltz, P. W., Weinberger, D. R., & Goldberg, T. E. (2007). Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia. *Schizophrenia Research*, 93(1–3), 304–316. <https://doi.org/10.1016/j.schres.2007.03.001>
- F20-F29 ICD-10 Version:2016. (n.d.).
- Foltz, P. W. (1996). Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments, and Computers*, 28(2), 197–202. <https://doi.org/10.3758/BF03204765>
- Harpe, S. E. (2015). How to analyze Likert and other rating scale data. *Currents in Pharmacy Teaching and Learning*, 7(6), 836–850. <https://doi.org/10.1016/j.cptl.2015.08.001>
- Howes, C., Lavelle, M., Healey, P. G., Hough, J. & McCabe, R (2017). Disfluencies in dialogues with patients with schizophrenia. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*. London, UK.
- He, Q. (2013). Text mining and IRT for psychiatric and psychological assessment. University of Twente.
- Iyer, D., Yoon, J., & Jurafsky, D. (2018). Automatic Detection of Incoherent Speech for Diagnosing Schizophrenia (pp. 136–146). *Association for Computational Linguistics (ACL)*. <https://doi.org/10.18653/v1/w18-0615>
- Just, S., Haegert, E. H., Kořánová, N., Bröcker, A.-L., Nenchev, I., Funcke, J., et al. (2019). Coherence models in schizophrenia. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*.
- Khokhlova, M. (2004). Building Russian Word Sketches as Models of Phrases, 364–371.
- Khudyakova, M. V., Bergelson, M. B., Akinina, Y. S., Iskra, E. V., Toldova, S., & Dragoy, O. V. (2016). Russian Clips: a Corpus of Narratives by Brain-

- Damaged Individuals. In *Proceedings of LREC 2016, Tenth International Conference on Language Resources and Evaluation*.
- Kraepelin, E. (1971). *Dementia Praecox and Paraphrenia*. Translated by RM Barclay.
- Kuperberg, G. R. (2010a). Language in Schizophrenia Part 2: What Can Psycholinguistics Bring to the Study of Schizophrenia and Vice Versa? *Linguistics and Language Compass*. <https://doi.org/10.1111/j.1749-818X.2010.00217.x>
- Kuperberg, G. R. (2010b). Language in Schizophrenia Part 1: An Introduction. *Linguistics and Language Compass*. <https://doi.org/10.1111/j.1749-818X.2010.00216.x>
- Lintean, M., Moldovan, C., Rus, V., & McNamara, D. (2010). The role of local and global weighting in assessing the semantic similarity of texts using latent semantic analysis. *Proceedings of the 23rd International Florida Artificial Intelligence Research Society Conference, FLAIRS-23, (Flairs)*, 235–240.
- Martin, E. A. Elizabeth A. (2010). *Concise medical dictionary*. Oxford University Press.
- Matell, M. S., & Jacoby, J. (n.d.). Is there an optimal number of alternatives for likert scale items? Study i: reliability and validity.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space, 1–12. <https://doi.org/10.1162/153244303322533223>
- Panicheva, P., & Litvinova, T. (n.d.). Semantic Coherence in Schizophrenia in Russian Written Texts.
- Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). Stroudsburg, PA, USA: Association for Computational Linguistics. <https://doi.org/10.3115/v1/D14-1162>
- Sharoff, S., & Khokhlova, M. (n.d.). Russian Web Corpus | Sketch Engine.
- Taiyun Wei, M. (2017). Visualization of a Correlation Matrix. *Statistician*, 56, 316–324.
- Tandon, R., Gaebel, W., Barch, D. M., Bustillo, J., Gur, R. E., Heckers, S., et al. (2013). Definition and description of schizophrenia in the DSM-5. *Schizophrenia Research*. <https://doi.org/10.1016/j.schres.2013.05.028>
- Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: a Professional Framework for Multimodality Research. In *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation*.
- World Health Organization (1992). *International classification of diseases and related health problems* (10th rev., ICD-10). Geneva, Switzerland: Author.



# 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, Cédric 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- $\beta$  (A $\beta$ ) plaques and tau tangles (Nelson et al., 2012). The recent

development of in vivo A $\beta$  and tau imaging confirms the hypothesis that A $\beta$  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 A $\beta$  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 A $\beta$  pathology as soon as possible, before the onset of memory impairment symptoms (McDade and Bateman, 2017).

However, detecting A $\beta$  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 A $\beta$  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 A $\beta$  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<sup>1</sup> (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.

---

<sup>1</sup>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.

When the study started, there were 102 subjects enrolled as controls and 204 as AD patients<sup>2</sup>. 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

---

<sup>2</sup>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<sup>3</sup>. 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

---

<sup>3</sup>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.

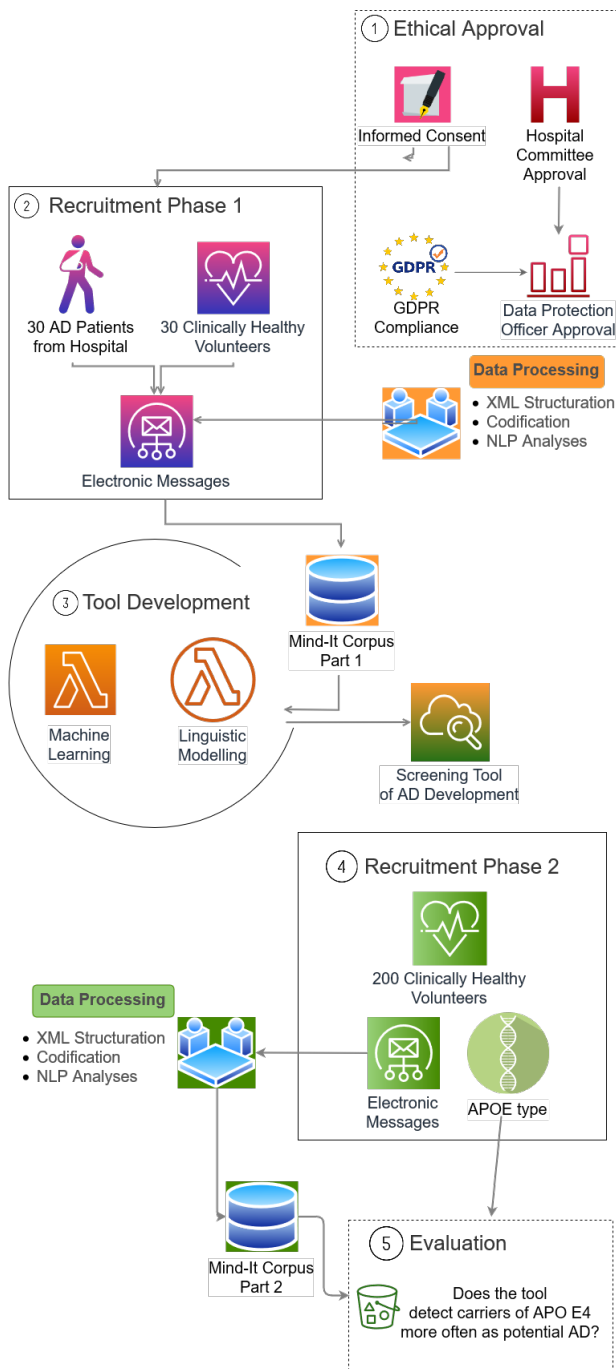


Figure 1: The phases of the Mind-It project.

### 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 puncture in which  $A\beta$  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 Coughon, 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  $\beta$ -amyloid and tauopathy. *JAMA Neurology*, 73(9):1070–1077.

# Coreference in Aphasic and non-Aphasic Spoken Discourse: Annotation Scheme and Preliminary Results

Svetlana Toldova<sup>1</sup>, Elizaveta Ivtushok, Kira Shulgina, Mariya Khudyakova<sup>2</sup>

<sup>1,2</sup>National Research University Higher School of Economics

21/4, Staraya Basmannaya str., Moscow, Russia

<sup>1</sup>toldova@yandex.ru, <sup>2</sup>mariya.kh@gmail.com

## Abstract

In this paper we propose the annotation scheme for the study of referential choice in spoken discourse of Russian speakers with and without aphasia. One of the key features of the annotation scheme is that it allows to establish not only coreference links between markables, but also annotate such phenomena as false-starts, repetitions and renamings. We analyzed the distribution of morphosyntactic types of referential expressions in narratives by people with fluent aphasia non-fluent aphasia and healthy speakers and found significant differences between the groups. We also discuss some important cases that were revealed by the qualitative analysis: use of zero and full anaphoric pronouns as introductory devices and inverse order of elements in noun phrases.

**Keywords:** coreference annotation, spoken corpora, aphasia, pear stories

## 1. Introduction

### 1.1 Referential Choice

Referential choice in language impairment has not been given much attention in the literature. For example, while many aspects of discourse production in aphasia were investigated in the last decades, mostly they do not specifically look at the referential options, focusing rather on cohesion errors (e.g. anaphoric pronouns without antecedents), lexical diversity or the proportions of content words, etc. (see Linnik, Bastiaanse and Höhle, 2016 for review). Studies focusing specifically on referential choice in aphasia discuss the frequency distribution of the basic NP types (full NP/anaphoric pronoun/zero pronoun) compared to the texts of healthy speakers (Romanova, 2010; Peng, 1992).

To track a referent in discourse, speakers can choose from a repertoire of referential expressions. They can use various descriptions that fully characterize a referent, such as “content” noun phrases (*a boy, the boy with the bicycle*), or, in case of repetitive mentions, they can use “reduced” expressions, such as demonstratives (*this, those*), or pronouns (*it, he*). One of the requirements regulating the choice is that the addressee should be able to recognize the coreferring expressions, or the expressions referring to the same entity. Establishing coreferential relations in discourse is a complex process depending on various cognitive, discourse and grammatical factors. This phenomenon is the topic of multidisciplinary research, unifying scientists in various fields of linguistics, such as syntax, computational linguistics, and psycholinguistics (Gordon and Hendrick, 1998).

It is a common assumption in cognitive modeling of referential choice that use of more or less linguistically reduced devices (such as nouns without modification or anaphoric pronouns) correlates with the cognitive status of a referent (cf. Ariel, 1991; Gundel et al., 1993; Kibrik et al., 2016; Prince, 1981). The referent status with respect to its

prominence, or topicality, in discourse imposes the constraints referential choice (e.g. the preference of anaphoric pronouns for more prominent referents, the “heaviest” NP for a first-time mentioned referent as *‘the boy who was riding a bike’*).

While the underlying cognitive mechanism of referential choice is assumed to be universal, different languages have different sets of possible referential expressions e.g. (Givón, 1983; Kibrik, 2009; Nedoluzhko et al. 2015; Romanova 2010). For example, in the so-called non-pro-drop languages (e.g. Germanic languages) omission of overt subject is ungrammatical, while in pro-drop languages zero pronouns are a valid referential option. As it has been shown in (Nedoluzhko et al., 2015), the distribution of zeroes and even anaphoric pronouns differs in English as a non-pro-drop language as compared to Russian that is a pro-drop one.

Referential choice can also depend on discourse conditions, e.g. different genres (Toole, 1996), discourse modality or individual speaker strategies (cf. Clancy, 1992).

In this paper we propose the annotation scheme for the study of referential choice in spoken discourse of Russian speakers with and without aphasia. We discuss the particular features of the material for annotation and the problems that it poses. We also provide preliminary results of quantitative and qualitative analysis of referential expressions in aphasic and non-aphasic discourse.

### 1.2 Coreference Annotation in Spoken Discourse

Coreference annotation, which is the focus of our research, has a relatively long history (cf. MUC-6 corpus<sup>1</sup>, Bagga & Baldwin, 1998). Some of the corpora with coreference annotation were created to evaluate automatic anaphora and coreference resolution systems (see for example the manuals for coreference annotation: Chinchor & Robinson, 1997; Hirschman & Thompson, 1997). Such corpora are also used in theoretical research focused on different

<sup>1</sup>[http://www-nlpir.nist.gov/related\\_projects/muc/proceedings/co\\_task.html](http://www-nlpir.nist.gov/related_projects/muc/proceedings/co_task.html)

features of referential choice (cf. Loukachevitch et al., 2011).

Depending on the purpose of the corpus, the annotation principles may vary. For example, the corpora of MUC-6 conference are focused on annotating only subclass of noun phrases, the ones referring to entities from the real world. In others, like ARRAU (Poesio & Artstein, 2008), the annotation of generic noun phrases is also presented. For Russian, the first open coreference corpus (RuCor<sup>2</sup>) was held in 2014. The majority of coreference corpora of written texts have clear-cut and well described annotation schemes.

Spoken discourse has some special features, such as unfinished utterances, various disfluencies, special interaction markers (see for example, Bergelson et al., 2015; Podlesskaya & Kibrik, 2007; Shriberg & Kwiatkowski, 1994). As the annotation procedure for this register is more complicated than for the written one, it needs further specification. There are numerous works on the annotation of various specific spoken discourse features. However, they focus primarily on the problems of discourse segmentation and different spoken discourse phenomena, such as hesitation pauses, self-corrections, discourse markers, markers of word-finding difficulties, and repetitions (MacWhinney, 2017; Podlesskaya & Kibrik 2007; Shriberg & Kwiatkowski, 1994; Varlokosta et al., 2016)

There are several corpora of spoken discourse annotated for coreference relations for some of the European languages. Some corpora consist of both spoken and written texts, such as the Polish Coreference Corpus (Ogrodniczuk et al., 2016), a parallel corpus of English and German texts, ParCor (Guillou et al., 2014), corpus of Dutch language COREA (Heindrickx et al., 2008) and a corpus of English texts of various genres, OntoNotes (Pradhan et al. 2007). Others contain only spoken discourse, for example, coreference corpus of the French language ANCOR\_Centre (Muzerelle et al., 2014).

The standard instructions for annotating written texts are not entirely suitable for annotating spontaneous speech (Kříž et al., 2015). One of the possibilities to override this problem is to have two separate annotation tiers: one for the transcribed discourse as is and the other one for the “normalized” or “reconstructed” text (Nedoluzhko et al., 2009). This would allow to analyze various disfluencies on a separate tier. This strategy has resulted in the majority of spoken corpora having coreference annotation only for the “normalized” level (e.g. ARRAU, Poesio and Artstein, 2008).

In our opinion, spoken discourse features, such as disfluencies, can affect various discourse mechanisms such as information flow manipulation and reference tracking. They can affect the choice of referential device or the assessment of referent’s discourse status (extra referent mentioning attracts more attention to it and thus influences the referent’s prominence assessment). All in all, we consider that deviations from “normalized” referential expressions and some types of disfluencies relate to the

naming procedure and should be integrated into the coreference annotation scheme. Thus, we propose the referential choice annotation scheme that includes annotation of different types of naming and reference mistakes, repetitions, and reformulations.

## 2. Material

Our experimental subcorpus is a part of the Russian CliPS corpus (Khudyakova et al., 2016) and contains narratives by people with two different types of aphasia (efferent motor and acoustic-mnemonic) and neurologically healthy speakers.

Aphasia is a language impairment resulting from damage in the language-dominant (usually left) hemisphere. Broadly aphasia can be divided into two types: with fluent and non-fluent speech output. Efferent motor aphasia is a representative example of non-fluent aphasia type. The word articulatory program breakdown is the main deficit that leads to inability to pronounce an organized set of articulation while producing a word (Akhutina, 2015). Another deficit in this type of aphasia appears on syntactic level and affects the syntactic schemata of sentences. It results in producing one-word utterances in severe forms of aphasia (“telegraphic speech”) or short noun-verb constructions in less severe cases. Nominative function is normally better preserved than predicative. Acoustic-mnemonic aphasia, belongs to the fluent aphasia type (Akhutina 2015). People with this type of aphasia experience auditory memory deficit which commonly leads to difficulties in remembering word sequences and sentences. It can also cause alienation of word meanings due to instability of the auditory images of words.

Basic annotation and segmentation into elementary discourse units (EDUs) was performed in ELAN (Wittenburg et al., 2006). Description of the Russian CliPS annotation scheme and characteristics of the speakers can be found in Khudyakova et al. (2016). The general statistics of the subcorpus for the current study are summarized in Table 1.

Parameter	Acoustic-mnemonic aphasia	Efferent motor aphasia	Healthy speakers	Total	
N texts	8	8	12	28	
Text length in tokens	Min	72	95	170	72
	Max	419	408	391	419
	Range	348	313	221	348
	Median	251	277.5	299	280.5
	Mean	263.3	245.6	277	264
Total	2106	1965	3324	7395	

Table 1. The general statistics of the subcorpus for the present study

For the present study we used two tiers of the annotation: quasi-phonetic, with markings for filled and empty pauses, laughter etc., and the lexical tier, the ‘normalized’ transcript with omission of pauses and fillers and standard orthography. Lexical transcripts were run through an automatic lemmatizer and morphological analyzer.

<sup>2</sup> <http://rucoref.maimbava.net>

### 3. Coreference Annotation Procedure and Tools

Considering the purposes of the study and the main features of the retellings in the corpora, the annotation procedure consists of several stages. First, an annotator needs to single out the referential expressions, or markables, that refer to the entities from the closed set. Second, each markable is assigned a morphosyntactic type. Next, we establish links between the markables. And finally, we annotate different types of errors. A summary of the annotation options is provided in Table 2, and an example of annotation is provided in Figure 1.

Annotation	Types	Subtypes
Morphosyntactic types	Full NPs	NPs with demonstratives
		NPs with other modifiers (adjectives, numerals, indefinite pronouns or a combination of them)
	Reduced referential devices	Anaphoric pronouns
		Relative pronouns
Links	Coreference	Coreference
		Non-coreferential relations
	Non-coreferential relations	False-start
		Repetition
		Self-correction
		Name elaboration
Errors	Morphological error	Lexical error
		Referential error

Table 2. Annotation options in the RuCor annotation scheme adapted for spoken discourse annotation

Figure 1. The coreferential links annotation tool

Coreference annotation was performed in a special coreference annotation tool that was designed for RuCor (Russian Coreference Corpus). It has a convenient web interface that allows parallel annotation by two or more people and online tracking of discrepancies between annotators. It also supports an extension of the feature set associated with a markable, and the linking of coreferent markables.

#### 3.1 Markables

##### 3.1.1 Set of Entities

Given the specific nature of the narratives in Russian CliPS, the set of referents mentioned in the stories is quite limited. Our decision was to choose a closed set of entities that are present in most of the retellings (people and objects appearing in the film). Thus, some of the referential expressions referring to other entities (for example, in some

speakers' comments or personal stories related to the film plot) were not analyzed.

We annotated markables of all types such as bare nouns, modified nouns (full noun phrases), anaphoric and relative pronouns, syntactic and non-syntactic zeroes. Zero pronouns were annotated as markables put in front of the corresponding verb, as in *Malchik padaet. Ø Ronyaet korzinu.* (The boy falls down. Ø Drops the basket'). Also non-referential entities' mentions, the NPs that are repetitions of referential NPs in false-start constructions, renaming constructions etc. (e.g. *jabloki, net grushi* 'apples, no, pears') were annotated as markables.

##### 3.1.2 Challenges in Markable Annotation

Many international standards for the annotation of written texts define a **markable** as a full noun phrase down to the nearest comma to the right (Krasavina, Chiarcos 2007). However, given the specific nature of spoken discourse, this principle as it is inapplicable in our corpus.

The first problem we are dealing with is various types of renaming. This construction is used when a false-name is followed by self-correction with a conjunction is quite frequent:

(1) *Malchik (ili paren)*  
'a boy (or a guy)'

Considering that both parts of such constructions are NPs with their own structure and lexical choice, these two NPs should be treated separately.

Secondly, we treat false starts as separate markables as well. By a false start we mean chunks of discourse where a participant begins to name an entity but stops or hesitates. Those are also annotated as separate NPs.

Another feature of spoken discourse is that an NP modifier (e.g. adjective or an apposition) can be postpositional:

(2) *... i proshli kak raz mimo khozyaina (grushy etoy bolshoi)*  
'lit. ...and passed by the owner (of the pear tree this big)' (c.f. this big pear tree)'

In written discourse, such postpositive adjective phrases are interpreted as parcellation or a detached phrase. There are no punctuation marks in spoken discourse, so there are certain difficulties in drawing clear-cut boundaries between phrases. We treat the postpositional adjective in spoken discourse as a part of the markable that includes the preceding head noun and its modifiers.

The third issue to be clarified is the referential properties of the markables. Though there are real-life entities appearing in the film such as trees in the garden or pears on these trees, the referential properties of corresponding NPs are not so clear in the retellings. A speaker can use an expression such as *sobiraet grushi* 'collects pears' in a generic sense. These entities are not "individualized" in the conceptualization of the corresponding scene by the speaker. We annotate this type of expressions as well.

### 3.2 Morphosyntactic Types of Markables

One of the main characteristics for a coreference phenomenon in a certain type of discourse (or a language) is the distribution of basic NP types.

We considered several types of **full NPs** in the process of annotation. The first type is a noun without modification which will be hereinafter referred as **bare noun**. A group of NPs consisting of **a noun with modification** contain several categories, among them NPs with adjectives (*shejny platok* ‘neckerchief’), numeral expressions (*tri korziny* ‘three baskets’), quantifiers (*neskolko grush* ‘several pears’), and demonstratives (*eta devochka* ‘this girl’) and their combinations (*dve korziny polnye* ‘two full baskets’). Standalone use of the modifiers is also considered as NP (*dlinnomu vtoromy* – lit. ‘to the long second’ meaning ... gave to the second boy who is tall).

As for reduced devices, we took into consideration different types of anaphoric expressions. First of all, we marked **anaphoric pronouns** (3<sup>rd</sup> person pronouns). Secondly, we marked **zero anaphoric pronouns**. Russian is a pro-drop language, meaning that a finite clause may have no overt subject. Even though the so-called zero pronouns can occur in written discourse, the nature of their use in spoken and written discourse is different. When using zero-type units in written text, one has to make sure the meaning of it is retraceable from the context. Nedoluzhko and colleagues (2015) claim that zero pronouns in Russian newswire texts appear only in syntactically motivated positions. However, this ‘syntactical motivation’ is not that crucial for spoken discourse: full NPs are mostly used when introducing the entity or removing it from the narration, while in between it is preferable to use pronouns and zeroes (Fox, 1987). The number of zero pronouns can be an important parameter in the analysis of pathological discourse compared to healthy discourse. Each predicate belongs to an elementary discourse unit (EDU), and we restore zero subjects for all the verb forms with no overt subjects (excluding syntactically motivated subject omission).

Another important type of referring expressions in both spoken and written discourse is the presence of **syntactic zeroes**: the absence of an overt noun phrase in a clause (e.g. ‘He took the basket and Ø drove away’). Thus, we marked the cases of subject co-ordination reduction (e.g. *...on vylozhl vse grushi i Ø polez opiat na grushu* – ‘... he laid out all pears and climbed again the pear (tree)’) as well as NP reduction in complement constructions as in *Malchik uронil korzinu kogda Ø padal* – ‘When falling down the boy dropped the basket’.

In addition, there are also **syntactic zeroes** in Russian that cannot possibly have an overt expression. The overt subject is impossible in an infinitival clause, in this case PRO (a pronominal determiner phrase that denotes an empty category, is used in non-finite clauses, and has strictly syntactic functions) is postulated (e.g. *I oni pomogli mal'chiku Ø sobirat' grushy s zemli*. – ‘And they helped the boy Ø to gather the pears from the ground’). We take these cases in consideration in our annotation scheme.

The use of syntactic and non-syntactic zeroes is a common strategy for coreferential choice in spoken discourse (Grenoble 2001). The number of zeroes used in retellings

may also be a crucial parameter when comparing discourse produced by people with aphasia and healthy speakers. The other two classes of syntactically motivated expressions that were annotated are reflexives and relative pronouns.

We also mark the demonstratives in NP position as well as the NP with the demonstratives as modifiers and relative pronouns. Though these types of NP are quite rare in our corpus, sometimes they occur in non-standard positions, namely they can be used as the first entity’s mention.

### 3.3 Links

Coreference relations were annotated with links between the markables referring to the same entities. In addition, we added several types of non-coreferential links, such as renaming, self-correction, false starts, alternative naming.

Those link types almost never appear in written discourse. However, these links are important for spoken discourse because they reflect the naming (or reference choice) procedure performed by a speaker as such. Thus, we have established a special feature, **link type**, in our annotation scheme. Our link taxonomy is based on the analysis of different types of non-coreferential expressions in spoken discourse provided in (Bergelson et al., 2015; Toldova et al., 2016). Consider the following examples:

- (3) *...eti samyie briuchki ... shtanishki*  
‘these trousers ... pants’ (renaming)
- (4) *... uvidel, chto v (odnoj) / (dve korziny), (dve korziny) polnyje grush*  
‘He saw that in one ... two baskets, two baskets are full of pears’
- (5) *a mimo (tri khuligana), (mozhet i nie khuligana)*  
‘beside three bullies, maybe not bullies’ (auto-correction)

Example (3) illustrates the renaming procedure where a speaker suggests a more precise common name for an entity. In (4) a speaker makes two false-starts. Firstly, he starts the locative phrase with a wrong numeral and then he changes the clause structure and the NP denoting the baskets becomes the subject. After then he repeats it. In (5) the process of auto-correction is verbalized. Firstly, a speaker suggests an NP ‘bullies’ to refer to the boys, then he uses appositive structure where he insert a modal *maybe* together with a negative particle to show that he is not sure that he has qualified the boys correctly. We consider these NPs in our annotation scheme since they may influence the level of a referent’s salience. Renaming and false starts are common in efferent motor aphasia.

One of the purposes of renaming strategy is to add a qualifying modifier in the second NP in renaming construction (6). Another one is a correction of referential expression choice, as in (7). Within the aphasic speakers group, the renaming is also used to auto-correct the morphological features of referring expressions, cf. (8) where singular is changed into plural in the second NP:

- (6) *Stoit (korzina), (korzina grush)*  
‘There is a basket, a basket with pears’
- (7) *khozyain etu, kazhduju grushu vytiral...*  
‘The owner wiped this, every pear’

- (8) *a fermer sobirajet (grushu), (grushi)*  
 ‘A farmer collects a pear, pears’

### 3.4 Errors

We introduced a special parameter for marking common errors when choosing a noun phrase. However, these errors require an additional, more thorough analysis. We classify the errors into three basic types: (a) morphological errors; (b) lexical choice; (c) referential choice. The errors are illustrated in (9), (10) and (11) respectively.

- (9) *pokazyvaiut sadovnika ... muzhchina kotoryi im ... na etogo malchika kotoryi*  
 ‘show a gardener ... a man who them ... on this boy who’ (number agreement in pronoun)

- (10) *paket nu...*  
 ‘bag well... (instead of basket)’

- (11) *on otblagodaril tremia grushami... eshcho chego... i vsio. on spuskaetsia vniz...*  
 ‘he<sub>i</sub> (the boy) thanked with pears... what else... and that’s it. He<sub>j</sub> (man) came down...’ (ambiguity)

There are morphological errors when an anaphoric pronoun chosen by the speaker disagrees in number or gender with its antecedent as in (9). In (10) the speaker has chosen a wrong lexeme for naming the referent: ‘bag’ instead of ‘basket’. Other cases of erroneous referential choice are cases, when speaker’s choice is ambiguous, an NP choice leads to a referential conflict (it can refer to more than one entity) or it can be mistakenly interpreted as referring to another entity. In (11) the speaker has chosen a wrong referential device, namely, he ambiguously uses the anaphoric pronoun ‘he’, it refers to the man though the most active referent in the previous sentence is ‘the boy’ (it is the subject of the previous sentence).

These error types pertain to the referential choice in spoken discourse and do not occur in written discourse.

## 4. Preliminary Analysis

### 4.1 Distribution of Main NP Types

The distribution of basic morpho-syntactic types in the subcorpus is presented in Table 3.

Chi-squared test showed a significant difference between the use of different NP types across all three groups of speakers ( $\chi^2 = 50.4203$ ,  $df = 12$ ,  $p\text{-value} = 1.179e-06$ ). However, the posthoc Kruskal-Wallis tests did not reveal significant differences between groups in the use of each morphosyntactic type except bare nouns ( $H = 6.3$ ,  $p = 0.04$ ).

### 4.2 Qualitative Analysis

As we can see the statistical tests show low difference in NP main types distribution among all the three groups. However, the qualitative analysis can reveal some important tendencies that would require further exploration on a bigger set of narratives.

Referential expression morphosyntactic type	Group		
	Acoustic-mnemonic aphasia	Efferent motor aphasia	Healthy speakers
Anaphoric pronouns	200 (25%)	154 (21%)	290 (25%)
Relative pronouns	18 (2%)	11 (1%)	18 (2%)
Reflexive pronouns	12 (2%)	5 (1%)	17 (2%)
Zero (+pro+PRO)	199 (25%)	162 (22%)	257 (22%)
Bare noun	250 (31%)	282 (38%)	344 (30%)
NP with a demonstrative	45 (6%)	20 (3%)	51 (4%)
Other NPs	69 (9%)	110 (15%)	190 (16%)
<b>Total</b>	<b>801</b>	<b>751</b>	<b>1162</b>

Table 3. Distribution of basic morphosyntactic types of referential expressions in the subcorpus

#### 4.2.1 Anaphoric expressions as introductory devices

It is generally assumed that using full NPs without any anaphoric expressions as introductory devices would be the only strategy that allows the speaker to avoid potential communicative failure. However, spoken register allows some departures from the assumed norm.

We analyzed the use of NP modifiers in interaction with the parameter of first / non-first entity mention and we have found some specific spoken discourse features in using anaphoric devices such as demonstratives in an introductory NP. It was already discussed that they are not normally used for introducing because of their status as the markers of high accessibility. Thus, in (12) the gardener is mentioned for the first time in discourse:

- (12) *... potomu chto etot chelovek, sobiraiushchii grushi, on naverniaka vsio-taki zvuk slyshit khorosho*  
 ‘because this man, who collects pears, he likely still hears sound well’

- (13) *oni nashli eje, shlyapu, kotoryje sletela, kogda on upal*  
 ‘They found it, the hat, that flew down when he fell down’

The possible explanation for this phenomenon is that the speaker assumes that the hearer is familiar with the referred entity as in (12). In this case the entity from the real world serves as the antecedent of the anaphoric device. Or the anaphoric pronoun is followed by a full NP as in (13).

#### 4.2.2 Zero anaphora as an introductory device

Another case of topicality hierarchy violation is the use of zeroes as introductory expressions. This case is rarer than the use of demonstratives, though we found several examples of it, cf. (14).

- (14)  $\emptyset$  *sobiral grushi, mimo projezzhal malchik na velosipede*  
 ‘ $\emptyset$  collected pears, a boy rode by on a bicycle’

### 4.2.3 Inverse order of elements in NP

Another interesting feature of spoken discourse is the inverse order of elements in a noun phrase, as in examples (15), (16), and (17). This word order is not ungrammatical in Russian, however it is not the default order.

(15) *uzhe gotovye v korzine sobrannye grushi*  
'already ready in a/the basket collected pears'

(16) ... *sobiral grushi krasivyye vkusnyye*  
'... collected pears beautiful tasty'

(17) ... *n'et korsiny odnoy*  
'... is absent basket one'

The phenomenon of NP modifiers postposition, while in Russian they normally located before the nucleus, can be caused by real time production of the narrative.

## 5. Conclusion

In this paper we described the annotation scheme adapted for coreference annotation of spoken discourse, including speech by people with aphasia.

A detailed qualitative analysis within different NP types and analysis of possible types of disfluencies concerning mentioning of different referents (false starts, renaming, self-corrections etc.) reveals some curious findings. Thus, in spoken discourse the anaphoric elements and even zeroes can be used for the first mention of an entity. The inverse order in NPs and the inclusion of epistemic modality expressions into them are possible. Although we did not get any prominent statistically significant differences between the speaker groups, we expect that further annotation and analysis of a larger sample of narratives will yield some important findings about referential choice in spoken discourse by people with and without aphasia.

## 6. Bibliographical References

- Akhutina, T. (2015). Luria's classification of aphasias and its theoretical basis. *Aphasiology*, 30(8), 878-897.
- Ariel, M. (1991). The function of accessibility in a theory of grammar. *Journal of Pragmatics*, 16(5), 443-463.
- Bagga, A. & Baldwin, B. (May 1998). Algorithms for scoring coreference chains. *The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, 1, 563-566.
- Bergelson, M. B., Akinina, Y. S., Dragoy, O. V., Iskra, E. V., & Khudyakova, M. V. (2015). Markers of word production difficulties in normal and clinical discourse production: continuity of norm in language and discourse. *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference Dialogue*, 14(1), 41-51.
- Chafe, W. L. (ed.). (1980). *The pear stories: Cognitive, cultural, and linguistic aspects of narrative production*.
- Chinchor, N. & Robinson, P. (1997). MUC-7 named entity task definition. *Proceedings of the 7th Conference on Message Understanding*, p. 29.
- Clancy, P. M. (1992). Referential strategies in the narratives of Japanese children. *Discourse Processes*, 15(4), 441-467.
- Cuenca, M. J. (2015). La connexió textual: l'adversitat en el nivell textual. *Caplletra. Revista Internacional de Filologia*, 7, 93-116.
- Fox, B. (1987). Anaphora in popular written English narratives. *Coherence and Grounding in Discourse*, 11, 157-174.
- Gordon, P. C. & Hendrick, R. (1998). The representation and processing of coreference in discourse. *Cognitive science*, 22(4), 389-424.
- Grenoble, L. A. (2001). Conceptual reference points, pronouns and conversational structure in Russian. *Glossos*, 1(1).
- Guillou L., Hardmeier C., Smith A., Tiedemann J. & Webber B. (2014) ParCor 1.0: A Parallel Pronoun-Coreference Corpus to Support Statistical MT, In *Proceedings of LREC 2014*, Reykjavik, Iceland
- Gundel, J. K., Hedberg, N. & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, 274-307.
- Heindrickx, I., Bouma, G., Coppens, F., Daelemans, W., Hoste, V., Kloosterman, G., Mineur, A.-M., Van Der Vloet, J. & Verschelde, J.-L. (2008). A coreference corpus and resolution system for Dutch. *Proceedings of LREC 2008*
- Hirschman, L. & Thompson, H. S. (1977) Overview of evaluation in speech and natural language processing. *Survey of State of the Art of Human Language*. 409-415.
- Khudyakova, M., Bergelson, M., Akinina, Y. & Iska, E. Russian CliPS: A Corpus of Narratives by Brain-Damaged Individuals. *Proceedings of LREC 2016 Workshop. Resources and Processing of Linguistic and Extra-Linguistic Data from People with Various Forms of Cognitive/Psychiatric Impairments (RaPID-2016)*.
- Kibrik, A. A. (2009) Basics of referential systems: Sorting things out. *Proceedings of the 7th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2009)*.
- Kibrik, A. A., Khudyakova, M. V., Dobrov, G. B., Linnik, A., & Zalmanov, D. A. (2016). Referential choice: Predictability and its limits. *Frontiers in Psychology*, 7, 1429.
- Kibrik, A. A., Podlesskaja, V. I. (2009). *Rasskazy o snovidenijah: korpusnoe issledovanie ustnogo russkogo diskursa*. Moscow, Jazyki Slav'anskoj Kul'tury.
- Krasavina, O. & Chiarcos, C. (2007). PoCoS-Potsdam coreference scheme. *Proceedings of the Linguistic Annotation Workshop*, 156-163.
- Kříž, V., Hladká, B. & Uřešová, Z. (2015) *Czech Legal Text Treebank*.
- Linnik, A., Bastiaanse, R. & Höhle, B. (2016). Discourse production in aphasia: A current review of theoretical and methodological challenges. *Aphasiology*, 30(7), 765-800.
- Loukachevitch, N.V., Dobrov, G.B., Kibrik, A.A., Khudyakova, M. & Linnik, A., 2011. Factors of referential choice: computational modeling. *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference Dialogue*, 10, 458-467.
- MacWhinney, B. (2017). Tools for analyzing talk part 1: The chat transcription format.
- Muzerelle J., Lefevre A., Schang E., Antoine J., Pelletier A., Maurel D., Eshkol I. & Villaneau J. (2014). ANCOR\_Centre, a large free spoken French coreference corpus: description of the resource and reliability measures. *Proceedings of LREC 2014*.

- Nedoluzhko, A., Mirovský, J., Ocelák, R. & Pergler, J. (2009). Extended coreference relations and bridging anaphora in the Prague Dependency Treebank. *Proceedings of the 7th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2009)*.
- Nedoluzhko, A., Toldova, S., Novák, M. (2015). Coreference Chains in Czech, English and Russian: Preliminary Findings. *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference Dialogue*, 14, 456-469.
- Ogrodniczuk M., Głowińska K., Kopeć M., Savary A., Zawisławska M. (2016) Polish Coreference Corpus. In: Vetulani Z., Uszkoreit H. & Kubis M. (eds) *Human Language Technology. Challenges for Computer Science and Linguistics*. LTC 2013. Lecture Notes in Computer Science, vol 9561.
- Peng, V. M. (1992). The usage of reference items in aphasic and normal conversations. *Journal of neurolinguistics*, 7(4), 295-307.
- Podlesskaya, V. & Kibrik, A. (2007). Samoisprialnenia govornjaš'ego i drugije tipy reč'evyh sbojev kak objekt annotirovanija v korpusah ustnoj reč'i. *Naučno-tehničeskaja informacija*, 2, 2-23.
- Poesio, M. & Artstein, R. (2008). Anaphoric annotation in the ARRAU corpus. *Proceedings of the 6th International Conference on Language Resources and Evaluation*. Marrakech.
- Pradhan, S. S., Ramshaw, L., Weischedel, R., MacBride, J. & Micciula, L. (2007). Unrestricted coreference: identifying entities and events in OntoNotes. *Proceedings of the International Conference on Semantic Computing (ICSC 2007)*, 446-453.
- Romanova A. Referential choice: Distribution of subject types in Russian aphasic speech. Presentation given at the Sixth Cambridge Postgraduate Conference in Language Research, Cambridge, UK.
- Shriberg, L. D. & Kwiatkowski, J. (1994). Developmental phonological disorders: a clinical profile. *Journal of Speech, Language, and Hearing Research*, 37(5), 1100-1126.
- Toldova, S. Y., Bergelson, M. B., & Khudyakova, M. V. (2016). Coreference in Russian Oral Movie Retellings (the Experience of Coreference Relations Annotation in "Russian CliPS" corpus ). In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016"* (Vol. 15, pp. 769–782).
- Toole, J. (1996). The effect of genre on referential choice. *Pragmatics and beyond new series*, 263-290.
- Varlokosta, S., Stamouli, S., Karasimos, A., Markopoulos, G., Kakavoulia, M., Nerantzini, M., ... & Protopapas, A. (2016). A Greek Corpus of Aphasic Discourse: collection, transcription, and annotation specifications. In *Proceedings of LREC 2016 Workshop. Resources and Processing of Linguistic and Extra-Linguistic Data from People with Various Forms of Cognitive/Psychiatric Impairments (RaPID-2016)*.
- Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: a professional framework for multimodality research. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, 1556-1559

## 7. Acknowledgements

This work was in part funded within the framework of a subsidy by the Russian Academic Excellence Project '5-100'.



# An NLP Pipeline as Assisted Transcription Tool for Speech Therapists

Gloria Gagliardi<sup>1</sup>, Lorenzo Gregori<sup>2</sup>, Andrea Amelio Ravelli<sup>2</sup>

<sup>1</sup>University of Naples “L’Orientale”, <sup>2</sup>University of Florence

gloria.gagliardi@gmail.com, lorenzo.gregori@unifi.it, andreaamelio.ravelli@unifi.it

## Abstract

This work presents the design of a computer-assisted transcription system for speech-language therapists and an evaluation of its core-module: the NLP pipeline. This pipeline combines a tokenizer, a lemmatizer, a part-of-speech tagger and a spellchecker to perform a semi-automatic annotation of speech transcriptions. The implemented module has been evaluated on a corpus of spoken interaction of children with Developmental Language Disorder (DLD) with the caregiver. Results are promising in automatic error detection (F-measure of 0.547 against a Ground Truth of 0.616) but low in automatic error correction, and confirm the effectiveness within an assisted transcription tool.

**Keywords:** Pathological Speech Processing, Developmental Language Disorder, PoS Tagging, Lemmatization

## 1. Introduction

Speech-language assessment and treatment are complex processes. Describing and interpreting children’s communication abilities entail the integration of a variety of information, gathered in the evaluation process (e.g. case history, review of sensory-motor and cognitive status, standardized and non-standardized measures of verbal and non-verbal language) (American Speech Language Hearing Association, 2004). The analysis of spontaneous and semi-spontaneous spoken productions of young patients is one of the essential elements for the formulation of logopedic balance, to ascertain the type, factor(s), and severity of the speech-language disorders (such as Speech Sound Disorder, Developmental Language Disorder and Social pragmatic Communication Disorder (American Psychiatric Association, 2013)), and to evaluate the expected habilitation or rehabilitation potential to set functional goals.

In the common practice, documentation of linguistic competence usually includes a portfolio of the child communication samples, e.g. transcript of audio or video-recorded interactions. To date, the collection and analysis of these data are very time consuming: as a matter of fact, Italian therapists manually transcribe the samples using phonetic alphabet (i.e. IPA, International Phonetic Alphabet), and this work is usually performed on “paper”. As a result, all the quantitative information which is needed for the evaluation (e.g. number/type of phonemic errors, number of tokens and lemmas, Mean Length of Utterance - MLU) is also empirically computed, representing a huge waste of time and resources.

### 1.1. Automatic annotation of pathological spoken language: a new challenge for the NLP community

Part-of-speech (PoS) tagging and lemmatization represent important preprocessing steps in Natural Language Processing: they are almost indispensable for the exploitation of corpus data and, since PoS tags are an essential input for most syntactic parsers, the accuracy of their annotation transitively worsens all the subsequent downstream higher level processing tasks (e.g. relation extraction) (Fan et al.,

2011; Ferraro et al., 2013).

POS-tagging is actually considered a solved task, since state-of-the-art taggers’ accuracy is around 97%–98% for English (Manning, 2011) and, nowadays, tools showing comparable outcomes are available for most languages, including Italian (Tamburini, 2007; Attardi and Simi, 2009; Tamburini, 2013).

As stated by (Giesbrecht and Evert, 2009) this means that, on average, every sentence contains a tagging error, but the accuracy of the system is close to the level of agreement between human annotators, and thus to the upper limit that can be expected from an automatic tool. This high accuracy is mostly attributable to the large amounts of tagged corpora, and the rapid progress in the study of corpus-based computational linguistics.

However, the state-of-the-art POS-taggers trained on written corpora do not provide satisfactory results if applied to spontaneous and semi-spontaneous spoken language (Uchimoto et al., 2002; Panunzi et al., 2004). Essentially, it is due to some peculiarities of the “oral medium”, namely freest word order, repetitions and fragmentation phenomena like false starts and interruptions.

Furthermore, PoS tagging and lemmatization tasks on speech corpora have not been tackled yet by the EVALITA periodic evaluation campaigns of NLP tools for the Italian language.<sup>1</sup>

Clearly, this lack in NLP for spoken Italian also affects the automatic analysis of children’s verbal productions and adult pathological language (e.g. aphasic speech).

The limited availability of data remains a stumbling block to reach state-of-the-art performances of NLP tools in the clinical domain. However, the number of computational applications is growing rapidly in the medical field: NLP techniques have been applied to the analysis of patients’ written and spoken texts, revealing latent patterns and regularities of their verbal productions, and thus acting as “digital biomarkers” (i.e. objective, quantifiable behavioral data which can be collected and measured through digital device, allowing for low-cost pathology detection and classification).

<sup>1</sup><http://www.evalita.it/>

## 2. Towards a computer-assisted transcription tool

Within the NLP tools for clinical application, we designed a system to support speech-language therapists in the error analysis of spoken productions. We aim at facing this issue by proposing an NLP pipeline for the assisted transcription and automatic analysis of speech recordings collected from Italian typical/atypical developing children. To the best of our knowledge, no previous study addressed this issue up till now for the Italian language.

In our intentions, the tool should support the speech-language therapists during all the phases, reducing their work burden. As a matter of fact, a simple but effective pipeline will allow the speech-language therapist to transcribe and automatically analyse spoken texts; the workflow can be summarised as follows:

1. *Transcription*: the user digitally transcribes the recorded samples, using the SAMPA phonetic alphabet (Wells, 1997).
2. *SAMPA to orthographic transcription converter*: the system converts phonetic transcriptions to regular Italian graphemes, so to be processed by an NLP pipeline.
3. *First automatic annotation*: tokenization, PoS Tagging and lemmatization of raw texts.
4. *Assisted transcription/correction module*: the system highlights “idiosyncratic words”, suggesting possible “corrections” by means of a spellchecker (e.g. *il lubo* > *il lupo*, en. *'the wolf'*).
5. *Manual correction of misspelled words*.
6. *Final automatic annotation*: PoS tagging and lemmatization of “normalized” texts.
7. *Statistics and IPA phonetic transcription generation*.

The full procedure requires limited user training. Italian therapists are usually reluctant to digitally transcribe, due to discomfort and concerns about the IPA keyboard. This difficulty can be easily overtaken using the SAMPA chart (Speech Assessment Methods Phonetic Alphabet), which is a machine-readable phonetic alphabet (Table 1).

As a matter of fact, the mapping of phonology into orthography is quite transparent and regular for Italian, where differences are limited to few phonemes. Therefore, phonetic and orthographic transcriptions are almost equivalent from a practical point of view. The initial effort is balanced out by the time saved in the analysis stage: after the final annotation, the system can quickly extract statistics at the phonological, lexical, and morpho-syntactic level, by comparing the raw transcription with the normalized one. For example, the following phonological processes can be easily identified:

- Consonant cluster reduction

[ˈkwɛsto] > [ˈkwɛtto] (*'this'*), [ˈskappa] > [ˈkappa] (*'runs away'*)

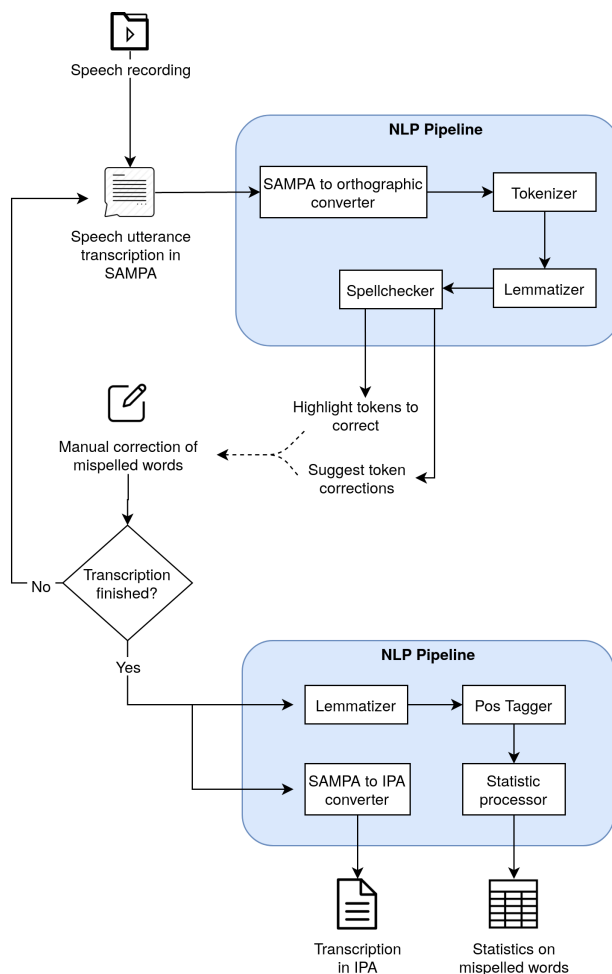


Figure 1: Full computer-assisted transcription pipeline.

Description	IPA	SAMPA
bilabial plosive	p b	p b
alveolar plosive	t d	t d
velar plosive	k g	k g
bilabial nasal	m	m
alveolar nasal	n	n
palatal nasal	ɲ	J
labio-dental fricative	f v	f v
alveolar fricative	s z	s z
palato-alveolar fricative	ʃ ʒ	S Z
alveolar affricate	ts dz	ts dz
palato-alveolar affricate	tʃ dʒ	tS dZ
alveolar trill	r	r
alveolar lateral	l	l
palatal lateral	ʎ	L
approximant	j w	j w
vowels	a e i ɛ ɔ u	a E e i O o u

Table 1: IPA and SAMPA phonetic alphabets.

- Consonant voicing

[ˈlupo] > [ˈlubo] (*'wolf'*)

Classical measures of lexical richness (e.g. Type/Token ratio) and syntactic development (e.g. MLU) can also be

automatically computed, lightening the workload. The pipeline can also generate an IPA transcription, which can be inserted in the patient’s portfolio, as requested by national good practice.

This paper presents the core module of the aforementioned pipeline (Figure 1), focusing on the ability to identify misspelled words, to suggest correction candidates and to automatically analyse transcriptions of pathological speech.

### 3. Material

To test the effectiveness of the pipeline, we rely on a small corpus of transcription of spontaneous speech interaction between infants and caregivers. This resource was designed to provide a first picture of narrative discourses produced by Italian monolingual preschoolers with Developmental Language Disorder (DLD) in comparison with typical peers matched by age.

DLD (previously known as Specific Language Impairment or SLI) is a neurodevelopmental disability which affects linguistic and communicative competence (American Psychiatric Association, 2013; Bishop et al., 2017): it is the most frequent developmental disorder in childhood, with an estimated overall prevalence in pre-school-aged children of about 7% (Tomblin et al., 1997; Johnson et al., 1999). It can selectively compromise all speech and language domains, affecting both language production and comprehension. A diagnosis of DLD should be stated (Bishop et al., 2017) for children showing a lower linguistic competence in comparison with the pairs; this verbal difficulty must affect patients’ everyday functioning and is unlikely to resolve by five years of age; in addition, it is not associated with a known cognitive, neurological or sensory-motor differentiating condition, depicting a more complex pattern of impairments, (e.g. brain injury, acquired epileptic aphasia in childhood, cerebral palsy, oral language limitations associated with sensorineural hearing loss as well as genetic conditions such as the Down syndrome).

To build our corpus, sixteen monolingual infants (13 M; 3 F) ranging in age from 4;2 to 5;4 (mean = 4;7) were enrolled. The sample was composed of a Control Group (CG) and a DLD Group, matched by age. The CG included eight participants (5 M; 3 F) without speech, language, hearing or cognitive impairments. The DLD group included eight male children who met the criteria for DLD with expressive deficits (American Psychiatric Association, 2013), recruited through the AUSL Toscana Centro. The diagnosis has been established according to national and international guidelines by expert clinicians, based on anamnestic data, clinical observation and standardized testing. Participants underwent a complete language evaluation, but particular attention has been paid to the assessment of children’s comprehension profile: all subjects performed within the normal range on the test of receptive vocabulary (TNL, Test Neuropsicologico Lessicale per l’età evolutiva (Cossu, 2013)), morpho-syntactic comprehension (TCGB, Test di Comprensione Grammaticale per Bambini (Chilosi and Cipriani, 2006) and PVCL, Prove di Valutazione della Comprensione Linguistica (Rustioni and Lancaster, 2007)) and listening comprehension (TOR, Test di Comprensione

del TestOrale 3-8 anni (Levorato and Roch, 2007)); therefore, expressive language problems occur essentially in isolation.

The corpus is composed by caregiver-child spontaneous speech interactions (duration: min. 3’51” - max. 23’53”), for a total of 1h57’41” transcribed audio-visual material. Oral production was elicited through three different tasks: the norm-referenced Bus Story Test (I-BST) (Renfrew, C.E., 2015; Cipriani et al., 2012; Mozzanica et al., 2016), and two semi-spontaneous retelling assessments, exploiting the renowned story Three Little Pigs (3LP), and a brand new short film called Little Polar Bear (LPB). While the I-BST examines story retelling with a colored picture support, the unnormed tests elicit children’s verbalizations through a paper book and a tablet respectively. During the 3LP task, children were asked to retell the renowned story using the pictures as prompts while flipping through the pages; in contrast, the LPB task was administered showing the video (around 100 seconds) to the child who was then requested to recount the plot while following the scrolling images without sound. None of the children knew the three stories. The trials were administered in a single test session of varying duration (~30 minutes).



Figure 2: The proposed tasks. From the left: “Bus Story Test”, “Three Little Pigs” and “Little Polar Bear”.

The tasks were recorded using a tablet placed in front of the subject. Data were transcribed using ELAN (Wittenburg et al., 2006).<sup>2</sup> Furthermore, transcriptions are also compliant with the L-AcT format (Cresti and Moneglia, 2018), a version of the standardized CHAT format (MacWhinney, 2000) enriched with the tagging of prosodic parsing. We chose the *utterance* as the reference unit in the speech continuum, defined as the counterpart of a speech act, namely ‘the minimal linguistic entity that can be pragmatically interpreted’ (Austin, 1962; Cresti and Moneglia, 2018)). Utterances are demarcated by prosody in the speech flow, therefore the identification of their boundaries is achieved through the detection of “prosodic breaks”. The identification of breaks reaches high inter-rater agreement in annotation, also among non-expert annotators (Cohen’s kappa for Italian around 0.8; (Danieli et al., 2004)), thus being a highly reliable chunking method.

### 4. The NLP pipeline in detail

The designed pipeline takes as input the text of the transcription of the session, and gives as output 2 objects: tran-

<sup>2</sup>All parents gave their consent to data recording, transcribing and processing.

scription in IPA characters, and detailed statistics of errors per part-of-speech and lemma.

Starting from the transcription of the session, the first step in the pipeline is the conversion from SAMPA to orthographic text, through a simple set of re-writing rules. Moreover, L-AcT specific tags and annotations (e.g. rephrasing, false starts, etc.) are removed, and the speaker’s turns are stored as distinct strings to be processed individually. In this way, it is possible to focus all the analysis exclusively on the child turns.

Then, each turn is tokenized and lemmatized with Tree-Tagger (Schmid, 1994), and all tokens are analysed by a spellcheck module. We used *pyspellchecker*,<sup>3</sup> a Python module that implements a Levenshtein Distance algorithm (Levenshtein, 1966) to find all possible permutations within an edit distance of 2 characters from each misspelled word. It then compares all permutations (character insertions, deletions, replacements, and transpositions) to known words in a word frequency list. As reference dictionary for the spellchecker module we used an Italian list of 50k words from the *WordFrequency Project*<sup>4</sup>, that has been extracted from the OpenSubtitles multilingual corpus<sup>5</sup> (Lison and Tiedemann, 2016).

The word that is found more often in the frequency list is more likely the correct result, and it is proposed as a substitution for the entry. At this point, the human annotator can chose to accept the proposed correction, or reject it and manually type the correct word. The index of the misspelled and its correction is stored, to be further used for the error analysis. The edited version of the text is then passed back to TreeTagger to perform lemmatization and POS-tagging. We perform lemmatization twice because misspelled words are initially tagged as “unknown”, and we make use of the tag shift in the error analysis.

Statistics on misspelled words can be easily obtained by parsing the annotated text. As an example, two fundamental pieces of information for a therapist are the set of wrong pronunciation of the same word and part-of-speech distribution during the speech.

Finally, the whole SAMPA transcript is converted to IPA, similarly to the very first step of the pipeline (SAMPA to orthographic), by following a simple set of re-writing rules, and the complete session is written out as a text file.

DLD Group			
	Child	CG	Total
<b>Tokens</b>	3367	3840	7207
<b>Words</b>	2191	2639	4830
<b>Unique words</b>	467	433	702
<b>Unique lemmas</b>	296	270	403
<b>Type/token ratio</b>	0.135	0.102	0.083

Table 2: Number of tokens, words and lemmas produced by children and care givers in the DLD Group sessions.

<sup>3</sup><https://pypi.org/project/pyspellchecker/>

<sup>4</sup><https://github.com/hermitdave/FrequencyWords>

<sup>5</sup><http://opus.nlpl.eu/OpenSubtitles2018.php>

Control Group			
	Child	CG	Total
<b>Tokens</b>	3419	2338	5757
<b>Words</b>	2345	1652	3997
<b>Unique words</b>	514	385	665
<b>Unique lemmas</b>	345	282	431
<b>Type/token ratio</b>	0.147	0.170	0.108

Table 3: Number of tokens, words and lemmas produced by children and care givers in the Control Group sessions.

#### 4.1. Corpus statistics

The automatic annotation through the NLP pipeline allowed us to derive some interesting information about the corpus used for this work. Tables 2 and 3 report the number of tokens, words, lemmas, and type/token ratio of DLD and Control Group sub-corpora. We can see that the number of words produced by children is similar between the two groups (2191 and 2345), attesting a substantial balance in the data.

An interesting result is that no relevant differences emerge between the two groups regarding the type/token ratio: 0.135 for DLD and 0.147 for Control. It derives that the speech of children with language disorder have roughly the same lexical variety than the speech of typical children. Otherwise, a big difference can be observed in caregiver speech (0.102 in DLD and 0.170 in Control), highlighting that caregivers talk is more simplified when addressed to children with language disorder.

PoS	DLD	Control
Noun	453 (20.68%)	453 (19.50%)
Verb	451 (20.58%)	457 (19.67%)
Conjunction	410 (18.71%)	384 (16.53%)
Article	256 (11.68%)	232 (9.99%)
Preposition	131 (5.98%)	139 (5.98%)
Adjective	108 (4.93%)	86 (3.70%)
Clitic	<b>90 (4.11%)</b>	<b>143 (6.16%)</b>
Adverb	89 (4.06%)	104 (4.48%)
Pronoun	<b>61 (2.78%)</b>	<b>105 (4.52%)</b>
Articulated Prep.	44 (2.01%)	66 (2.84%)
Determiner	31 (1.41%)	48 (2.07%)
Auxiliary verb	23 (1.05%)	28 (1.21%)
Negation	20 (0.91%)	39 (1.68%)
Word “che”	17 (0.78%)	28 (1.21%)
WH Word	5 (0.23%)	2 (0.09%)
Proper Noun	2 (0.09%)	4 (0.17%)
Number	0 (0.00%)	5 (0.22%)

Table 4: Part-of-speech distribution in children speech (in DLD and Control Groups).

Finally, when looking at the part-of-speech distribution of children in the two groups (Table 4) we could not find huge differences, but a notable gap can be observed in the production of clitics and pronouns, where numbers are lower in DLD Group ( $\chi$ -squared test with  $p$ -value < 0.001). This

seems to confirm and enrich known data about clitic productions in Italian impaired children (Bortolini et al., 2006; Guasti et al., 2016), even if further analyses are needed to support this argument.

## 5. Evaluation

Table 5 reports the output of the error analysis performed within the NLP pipeline. POS\_unknown refers to the lemmas not recognized by the POS-tagger, while Spellcheck stands for the words reported by the spellchecker. It is possible to notice a slight difference between the phenomena highlighted by the two methods.

	DLD	Control
<b>POS_unknown</b>	90 (4.11%)	40 (1.71%)
<b>Spellcheck</b>	84 (3.83%)	55 (2.35%)

Table 5: Number of words tagged as “unknown” by the POS-tagger and marked by the spellchecker. Percentages are reported with respect to the total number of words in each group.

To evaluate the results of error identification and automatic correction tasks, we built a gold standard through manual annotation of the children turns in the whole corpus. Each misspelled word were marked and annotated with the correct version. Data reported in Table 6 show that, as expected, the DLD Group has a double rate of misspelled words than the Control Group: 4.11% of the total produced words in DLD are misspelled against 2% in Control. Moreover, it is important to highlight that a significant number of misspelled words are not recognized by the human annotator which marked them with “unknown” during the manual check. In total, there are 8 words in the DLD Group and 10 in the Control group, for a total of 13,14% of misspelled words.

	DLD	Control
<b>Manual corr. (MC)</b>	90 (4.11% w.)	47 (2.00%)
<b>MC unclassified</b>	8 (8.89% MC)	10 (21.28%)

Table 6: Number of manual corrections in the gold standard (total and unknown words).

Accuracy in both error detection and correction is reported in Table 7. For the error detection task we reported the number of manual corrections matching with the “unknown” tag of the lemmatizer (Lem) and the number of manual corrections matching the words marked by the spellchecker (SC). Automatic correction task is performed by the spellchecker only and the numbers regard automatic corrections matching with manual corrections. Precision, Recall and F-measure are computed for both tasks and reported in Table 8.

It is important to highlight that the proposed system is not able to identify any of the cases in which the misspelled word is still a word form that exists in the language. These cases are frequent in Italian, especially with short words, like articles or prepositions, where it is likely that deletion

	DLD	Control
<b>Err. detection (Lem)</b>	55 (61.11%)	18 (38.30%)
<b>Err. detection (SC)</b>	48 (53.33%)	18 (38.30%)
<b>Err. correction (SC)</b>	27 (32.14%)	4 (7.27%)

Table 7: Numbers and percentages of misspelled words properly detected by lemmatizer (Lem) and spellchecker (SC), and properly corrected ones by the spellchecker.

	Pr	Rec	Fm
<b>Error detection (Lem)</b>	0.562	0.533	0.547
<b>Error detection (SC)</b>	0.475	0.482	0.478
<b>Error detection (GT)</b>	0.681	0.562	0.616
<b>Error correction (SC)</b>	0.223	0.304	0.257

Table 8: Precision, Recall and F-measure of the error detection task for lemmatizer (Lem), spellchecker (SC) and Ground Truth (GT), and of the error correction task for spellchecker.

or substitution of a single phoneme produce a proper word (e.g. *il* > *i*; *del* > *dei*). For this reason, the maximum Recall that our system can reach in error identification task cannot be very high: with the given dataset, considering only errors that produce impossible words, we obtained a ground truth Recall of 0.562. Table 8 shows that there is a low margin of improvement. Conversely, the Precision of the system is deeply affected by those lexical productions that are specific of spoken language, like interjections, vocalizations and filled pauses (e.g. *ehh*, *mah*, *mmm*), which are wrongly marked as errors. Considering these expressions in our dataset as constrained false positive, we obtained a ground truth Precision of 0.681.

While a substantial Recall improvement is not possible with the given system - because it would require additional NLP modules of language understanding - Precision in error detection could be improved a lot, by upgrading the pipeline with NLP tools (spellchecker and lemmatizer) trained on spoken corpora.

As stated before, some errors cannot be satisfactorily managed by the pipeline. As an example, there are some phonological processes that are typical in children linguistic development which result in real words (e.g. [‘tʃuffo] > [‘tuffo], *stopping*, en. ‘lock of hair’ > ‘dive’; [ba‘nana] > [‘nana], *weak syllable deletion*, en. ‘banana’ > ‘dwarf’) and neologisms like ‘peciano’, ‘selfia’ or the portmanteau ‘fangua’ (coined by blending ‘fango’ and ‘acqua’, en. ‘mud’-‘water’). These phenomena are not understandable by the therapist outside their linguistic and extra-linguistic contexts. On the contrary, simple heuristics can be incorporated into the pipeline to manage high-frequency articulation or phonological error patterns that characterised typical and atypical developmental trajectories. For example, the already mentioned cluster reduction (e.g. [‘kwesto] > [‘kwetto], en. ‘this’, or [‘skappa] > [‘kappa], en. ‘ø run away’), prevocalic consonant voicing (e.g. [‘lupo] > [‘lubo], en. ‘wolf’) or deaffrication ([‘gottʃe] > [‘gosse], en. ‘drops’).

By considering these data and their analysis, we can derive that a semi-automatic system of computer-assisted translation, as proposed in this paper, appears to be more suitable than a fully automatic one, that provides an automatic annotation of transcripts. In fact, results of error detection are promising and can be fruitfully exploited to highlight misspelled words, while accuracy on automatic correction is low and definitely not reliable to replace manual annotation. However, proposed corrections can be very useful to save time during the transcription, showing a set of possible correction options that can be selected by the annotator. To this aim, a simple caching system of annotated data would bring a strong improvement to the spellchecker, given that, phonological errors tend to be recurrent: in our test corpus, 40.33% of the pairs [correct word, misspelled word] occur more than once.

## 6. Conclusions and future work

This work discussed the application of an NLP pipeline within a computer-assisted transcription system. The system architecture foresees a SAMPA transcription of pathological speech and aims at helping speech therapists to annotate misspelled words, to produce useful statistics on errors in words production, and to generate text in IPA. The core module of the system was developed and analyzed through a spoken corpus of children with Developmental Language Disorder. The tasks considered are automatic detection and automatic correction of misspelled words. The evaluation highlights an average accuracy on error detection and a low accuracy on error correction. However the results appear to be relevant for the proposed application. It is important to notice that a naive spellchecker module was implemented, thus more sophisticated systems may be able to improve also error correction results. It is important to point out that the lack of large annotated speech corpora for Italian (and in particular for first language acquisition) is the main obstacle to a more effective system. In fact, many of the problems highlighted in this paper would be correctly handled by NLP tools specifically trained on spoken Italian. The presented analysis represents the first step in the construction of a full transcription tool that will be developed as an editor for speech therapists (in the form of a standalone software or ELAN plugin).

## 7. Acknowledgements

The authors are deeply grateful to Francesca Beraldi and Milvia Innocenti who collected the clinical data. The precious help of Annalisa Raffone is also acknowledged.

## 8. Author contribution

As far as academic requirements are concerned, sections 1 and 3 were authored by Gloria Gagliardi, who also performed the manual annotation for the pipeline evaluation; Lorenzo Gregori wrote sections 2 and 5 and computed the statistics; Andrea Amelio Ravelli takes official responsibility for section 4 and implemented the NLP pipeline. All authors read and gave final approval for submission. The usual disclaimers apply.

## 9. Bibliographical References

- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5*. American Psychiatric Association, Washington, DC.
- American Speech Language Hearing Association. (2004). *Preferred Practice Patterns for the Profession of Speech-Language Pathology*.
- Attardi, G. and Simi, M. (2009). Overview of the evalita 2009 part-of-speech tagging task. In *Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence, 12th December 2009, Reggio Emilia, Italy*.
- Austin, J. (1962). *How to Do Things with Words*. Clarendon Press, Oxford.
- Bishop, D., Snowling, M., Thompson, P., Greenhalgh, T., and THE CATALISE-2 CONSORTIUM. (2017). Phase 2 of catalise: a multinational and multidisciplinary delphi consensus study of problems with language development: Terminology. *Journal of Child Psychology and Psychiatry*, 58(10):1068–1080.
- Bortolini, U., Arfé, B., Caselli, C. M., Degasperi, L., Deevy, P., and Leonard, L. B. (2006). Clinical markers for specific language impairment in italian: the contribution of clitics and non-word repetition. *International Journal of Language & Communication Disorders*, 41(6):695–712.
- Chilosi, A. and Cipriani, P. (2006). *TCGB. Test di comprensione grammaticale per bambini*. Edizioni Del Cerro, Tirrenia.
- Cipriani, P., Salvadorini, R., and Zarmati, G. (2012). *Bus Story Test. Test di valutazione delle abilità narrative*. Edizioni La Favelliana, Milano.
- Cossu, G. (2013). *TNL. Test neuropsicologico lessicale per l'età evolutiva*. Hogrefe Editore, Firenze.
- Cresti, E. and Moneglia, M. (2018). Chapter 13. the illocutionary basis of information structure: The language into act theory (I-act). pages 360–402.
- Danieli, M., Garrido, M., Moneglia, M., Panizza, A., Quazza, S., and Swerts, M. (2004). Evaluation of Consensus on the Annotation of Prosodic Breaks in the Romance Corpus of Spontaneous Speech “C-ORAL-ROM”. In Maria Teresa Lino, et al., editors, *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, pages 1513–1516, Paris. ELRA.
- Fan, J.-W., Prasad, R., Yabut, R. M., Loomis, R. M., Zisook, D., Mattison, J. E., and Huang, Y. (2011). Part-of-speech tagging for clinical text: wall or bridge between institutions? *AMIA Annual Symposium*, pages 382–91.
- Ferraro, J. P., Daumé, H., DuVall, S. L., Chapman, W. W., Harkema, H., and Haug, P. J. (2013). Improving performance of natural language processing part-of-speech tagging on clinical narratives through domain adaptation. *Journal of the American Medical Informatics Association: JAMIA*, 20(5):931–9.
- Giesbrecht, E. and Evert, S. (2009). Is part-of-speech tagging a solved task? an evaluation of pos taggers for the

- german web as corpus. In I. Alegria, et al., editors, *Proceedings of the 5th Web as Corpus Workshop (WAC5)*.
- Guasti, M. T., Palma, S., Genovese, E., Stagi, P., Saladini, G., and Arosio, F. (2016). The production of direct object clitics in pre-school- and primary school-aged children with specific language impairments. *Clinical Linguistics & Phonetics*, 30(9):663–678. PMID: 27285056.
- Johnson, C., Beitchman, J., Young, A., Escobar, M., Atkinson, L., Wilson, B., Brownlie, E., Douglas, L., Taback, N., Lam, I., and Wang, M. (1999). Fourteen-year follow-up of children with and without speech/language impairments: speech/language stability and outcomes. *Journal of Speech, Language, and Hearing Research*, 42(3):744–760.
- Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710.
- Levorato, M. and Roch, M. (2007). *TOR. Test di Comprensione del Testo Orale - 3-8 anni*. Giunti O.S., Firenze.
- Lison, P. and Tiedemann, J. (2016). Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.
- MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk*. Lawrence Erlbaum Associates, Mahwah.
- Manning, C. D. (2011). Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In *Proceedings of the 12th international conference on Computational linguistics and intelligent text processing - Volume Part I, CICLing'11*, pages 171–189. Springer-Verlag.
- Mozzanica, F., Salvadorini, R., Sai, E., Pozzoli, U., Maruzzi, P., Scarponi, L., Barillari, M., Spada, E., Ambrogi, F., and Schindler, A. (2016). Reliability, validity and normative data of the italian version of the bus story test. *International Journal of Pediatric Otorhinolaryngology*, 89:17–24.
- Panunzi, A., Picchi, E., and Moneglia, M. (2004). Using pitagger for lemmatization and pos tagging of a spontaneous speech corpus: C-oral-rom italian. pages 563–566. ELRA, Paris.
- Renfrew, C.E. (2015). *Bus Story Test. A test of narrative speech*. Speechmark, Milton Keynes (UK).
- Rustioni, D. and Lancaster, M. (2007). *PVCL Valutazione della Comprensione Linguistica*. Giunti O.S., Firenze.
- Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.
- Tamburini, F. (2007). Evalita 2007: The part-of-speech tagging task. *Intelligenza Artificiale*, IV(2):4–7.
- Tamburini, F. (2013). The lemmatisation task at the evalita 2011 evaluation campaign. In B. Magnini, et al., editors, *Evaluation of Natural Language and Speech Tools for Italian*, pages 230–238. Springer.
- Tomblin, J., Smith, E., and Zhang, X. (1997). Epidemiology of specific language impairment: Prenatal and perinatal risk factors. *Journal of Communication Disorders*, 30(4):325 – 344.
- Uchimoto, K., Nobata, C., Yamada, A., Sekine, S., and Isahara, H. (2002). Morphological analysis of the spontaneous speech corpus. In *COLING 2002: The 17th International Conference on Computational Linguistics: Project Notes*.
- Wells, J. C. (1997). Sampa computer readable phonetic alphabet. *Handbook of standards and resources for spoken language systems*, 4.
- Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). Elan: a professional framework for multimodality research. pages 1556–1559.



# An Exploration of Personality Traits Detection in a Spanish Twitter Corpus

Gerardo Sierra<sup>1</sup>, Gemma Bel-Enguix<sup>1</sup>, Alejandro Osornio-Arteaga<sup>1</sup>, Adriana Cabrera-Mora<sup>1</sup>,  
Luis García-Nieto<sup>1</sup>, Alfredo Bustos<sup>2</sup>, Ana-Miriam Romo-Anaya<sup>2</sup>, Víctor Silva-Cuevas<sup>2</sup>

Universidad Nacional Autónoma de México

Ciudad de México

Instituto Nacional de Estadística y Geografía

Aguascalientes

{gsierram, gbele, aosornioa, acabreram, lgarciani}@ingen.unam.mx

{Alfredo.Bustos, Miriam.Romo, VictorSilvaC}@inegi.org.mx

## Abstract

The article explores the identification of personality traits in a Twitter corpus in Spanish. Having this objective, we use a dictionary oriented program named LIWC, that support the study of some linguistic characteristics, especially lexical ones. Due to the lexical approach of this program, we argue the importance of extending the basic dictionaries of the program in order to include the Social Media slang, netspeak. In this way, information from internet texts can be better captured with the purpose of improving a psychological analysis. Following the Big Five Personality Model, a clustering analysis of 98 Twitter users was carried out considering previous significant correlations between word categories and personality traits. Five groups were identified and we discuss the likely representation of the personality traits by each one of them. Finally, both the contributions and limitations of the present study are commented upon, as same as the challenges to still meet in this research area.

**Keywords:** psychological analysis, netspeak, dictionary

## 1. Introduction

The internet is the most comprehensive and accessible source of texts. This variety of texts includes people's online searches and their daily posts on social networking websites, in which they share about their interests, thoughts or any topic of their choice. Perhaps because of the large amount of time that internet users spend online, it has become possible to take a look at the world of these individuals through their online behavior and specially, through their language. Additionally, since these internet texts are being produced freely and without any constrained context, they represent a source of information that reflects a naturalistic production of language (Boyd and Pennebaker, 2015).

For many disciplines it has become very important to take a deeper look into these online spaces. Psychology has been, since its origins, closely linked with the study of language (Miller, 1990). In therapy, the importance of a person's life story has been always taken intuitively. However, the amount and choice of certain words have been taken only till recently into consideration. For this reason, the study of language use in social media presents a good opportunity to expand our knowledge about the relationship between word usage and psychological traits.

Researchers have been already able to find the value of individual words as psychological data, making possible to draw inferences about the individuals that produce those words. This possibility has been specially true for research that focuses on the relationships between personality traits and word usage (Ireland and Mehl, 2014). One of the strongest arguments in favor of this position is provided by James Pennebaker, an active researcher in this field: "The words we use in daily life reflect who we are and the social relationships we are in... Language is the most common and reliable way for people to translate their internal thoughts and emotions into a form that others can under-

stand. Words and language, then, are the very stuff of psychology and communication" (Tausczik and Pennebaker, 2010, p.25).

Due both to the linkage between language and personality and the rise of the internet as a valuable source of naturally produced language data, it has become possible to identify certain personality traits by analyzing the language that people display on the internet. Following the Big Five personality model (Goldberg, 1990), which has been found to correlate with linguistic features (Yarkoni, 2010; Qiu et al., 2012; Hirsh and Peterson, 2009), this paper is a preliminary approach to the detection of personality traits on Twitter with a Mexican Spanish corpus.

For this work, we used the program Linguistic Inquiry and Word Count (LIWC) (Pennebaker and Francis, 1996). In Section 2 we mention some previous work on personality and language together with LIWC and its features. Section 3 analyzes the main linguistic features of Netspeak, i.e. the specific slang used in Social Media. Some ideas about how Netspeak can be integrated into psychological studies are provided in Section 4. In Section 5 we explain the corpus and methodology, while the results are presented in Section 6 and are discussed in Section 7. The paper closes with Section 8 devoted to the present study limitations and future steps in this incipient line of research.

## 2. Related Work

Psychology's relationship with language has always been a close one. Much work has been done regarding how psychological phenomena are reflected through the use of language. Some examples can be found in social status relationship dynamics (Kacewicz et al., 2013), marked language patterns which indicate when someone may be lying (Hancock et al., 2008), or in the subtle language variations between people with mental disorders and people without them (Coppersmith et al., 2015). Another special interest for psychological studies is the one about personality. The

question about the manifestation of personality through language has relied on one of the most acknowledged personality theories, the Big Five personality traits, in order to analyze how it relates to natural language use. This approach to personality is mainly lexical in nature, focusing on the meanings of the words that people use to describe others and themselves, thus, natural language must be studied with the aim of making accurate descriptions of personality (Goldberg, 1981). The Big Five model states that personality's universal structure is manifested through five broad dimensions, these being: Neuroticism, Extraversion, Agreeableness, Conscientiousness and Openness to Experience.

The trait Neuroticism reflects a tendency to experience psychological distress; Extraversion reflects mainly social traits and experiences of positive emotions; Agreeableness points to a dimension of interpersonal behavior, reflecting cooperation and trust-like traits; Conscientiousness reflects traits of scrupulous and planning individuals; and Openness to Experience refers to imaginative, intellectual and flexible-style traits (Costa and McCrae, 1992). Given the lexical nature of this approach to personality, it comes as no surprise that a good amount of research regarding this theory has been conducted using natural language as valid data to establish relationships with all five personality dimensions. One of the first and most important studies that follows this approach was conducted by Pennebaker and King (1999). In this study, the authors successfully correlated the language use of what they called linguistic styles in various texts from students with personality measures from the Big Five personality traits. These linguistic styles were made up of the functional words that the participants used in their texts and represent a different approach to content words. Findings report that functional words use can unveil individual differences, specially in regards to personality. Similarly, a study by Hirsh and Peterson (2009) analyzed the language of undergraduate students in an assigned self-narrative written task and correlations were made between word usage and Big Five personality scores. The results showed that word choice was significantly associated with Big Five personality traits across psychological categories. Both of these studies handled their respective data with a special program called LIWC (Linguistic Inquiry and Word Count) (Pennebaker and Francis, 1996), which is a program that takes on the task of analyzing language in a more psychological way.

LIWC consists of a word counter and a large internal dictionary with several words categorized into psychologically relevant categories. Its psychometric validity has been demonstrated (Pennebaker et al., 2015). The program operates by identifying the words in a text file that corresponds with the dictionary and assigning them to the categories. As a result, the program calculates the frequency of each word category in relation to the total words of the uploaded text. Thus, scores of each word category are shown in the output. Such data allow the establishment of relationships with numerous topics via word usage. Moreover, there are also validated versions in different languages, such as Spanish (Ramírez-Esparza et al., 2007), which facilitates cross-cultural research with the program.

Research using LIWC, since the program's creation until now, and in conjunction with other measures, has found one of its most successful applications in the field of personality expression through language. Moreover, the jump made by written language to the internet has given rise to psychological studies that work with texts found in online spaces such as social media. These social environments on internet, such as social media, blogs, emails, etc., give the opportunity to analyze productions of language and its linkage to personality. For example, Qiu et al. (2012) studied the language used by users of the social media site Twitter using LIWC. Word categories scores were correlated with personality scores from a Big Five Personality inventory. Naturally, the authors found correlations between words used in tweets and the Big Five personality traits. Another similar study was carried out by Yarkoni (2010), who collected large samples of text from 694 internet blogs, containing around 100,000 words each. Again, scores from word categories and personality traits were correlated for each of those blogs' authors. Highly valuable data were obtained and it was possible to identify interesting relationships between online exhibited language and personality, this being consistent with reported relationships between LIWC's word categories, personality scores and language usage across multiple studies (see Table 1). Because the Yarkoni (2010) study analyzed large amounts of text data for each user, the present study includes those categories with significant correlations into the analysis (see Section 5.2.2).

All of these studies have relied on LIWC and its word categories in order to handle their language data. However, the program presents some limitations that are worth mentioning here. As a word counting program, it takes each individual word in a text regardless of its context. Thus, in a more intuitive sense, it hinders the value of the results in some way. As it will be discussed below, this must not be entirely the case, because this apparent limitation can be reinterpreted to benefit the value of the results given by the program. Another limitation of LIWC is addressed by Schwartz et al. (2013). The authors' argument is that LIWC's analytic power highly depends on how extensive its dictionary is, thus following a "closed" approach to language analysis. Therefore, the authors propose an alternative to LIWC's usage in the form of an "open vocabulary approach" which, unsurprisingly, was also applied to personality and language. This open vocabulary approach includes phrases and individual words along with automatically generated topics rather than an priori-made lexicon, all incorporated into a model, and presented in a word cloud format accompanied by word, phrase and topic correlations to personality scores of Facebook users and their profile updates as the analyzed language data. The authors found good values and correlations with all five personality dimensions of the Big Five, some even outperforming the ones found by LIWC-reliant approaches.

Similarly, Majumder et al. (2017) approached the Big Five personality traits and language using a non-LIWC approach. They extracted personality traits from essays using a convolutional neural network method. For this, they trained five different networks, one for each personality

Neuroticism	1st person singular pronoun, 2nd person pronoun (-), negation, article (-), negative emotion, anxiety, anger, cognitive process, causation, discrepancy, tentative, certainty, feeling, friend (-), space (-), exclusive, swear
Extroversion	1st person plural pronoun, 2nd person pronoun, number (-), positive emotion, causation (-), inhibition (-), tentative (-), certainty, sensory, hearing, social process, friend, family, human, inclusive, work (-), achievement (-), music, religion, physical state, sexuality
Openness	Pronoun (-), 1st person singular pronoun (-), 1st person plural pronoun (-), 1st person pronoun(-), 2nd person pronoun (-), negation (-), assent (-), article, preposition, number (-), affect (-), positive emotion (-), cognitive process (-), discrepancy (-), sensory (-), social process (-), family (-), human (-), time (-), past tense verb (-), present tense verb (-), future tense verb
Agreeableness	Pronoun, 1st person plural pronoun, 1st person pronoun, numbers, positive emotion, negative emotion (-), anger (-), causation (-), seeing, feeling, social process, friend, family, time, past tense, verb, space, inclusive, motion, leisure, home, death (-), physical state, body state, sexuality, swear (-)
Conscientiousness	Negation (-), negative emotion (-), anger (-), sadness (-), cognitive process (-), causation(-), discrepancy (-), tentative (-), certainty (-), sensory process (-), hearing (-), human (-), time, exclusive (-), achievement, death (-), swear (-)

Table 1: Correlations between Big Five personality traits and LIWC Categories (Yarkoni, 2010)

trait and made them binary classifier for the corresponding trait to be positive or negative, additionally, using convolution filters they gave sentences from the essays to obtain a sentence model that resulted in the form of n-gram feature vectors. This approach is a deep learning document modeling one, and relies heavily on how well the network is trained for reliable results. The authors obtained good results in all five dimensions of the Big Five personality traits. The alternatives to LIWC listed here each have their own

advantages and disadvantages as any research method carries, however, and despite outperforming LIWC's execution in some areas, the word counting program still has some strong points in its favor that justify using it over these other approaches. One argument in favor of LIWC's usage comes from Ireland and Mehl (2014) who argue that despite the program's inability to detect context in text, it can be used to examine the focus that people place on certain words and types of words and how their presence signals the importance that these have over others and how they relate to the overall subject of their texts (for further clarification, revise the cited reference). Another of LIWC's advantages is that its creation serve psychological-oriented purposes. Therefore, this program is oriented towards research of this kind, on top of it being the preferred resource for analysis of this kind and possessing a wide range of studies that correlate its categories with various topics, giving researchers a bigger pool of results that share a common ground and allow for more comparisons to be made between them. These arguments motivate the decision to use LIWC and its categories in this research.

With this new direction of psychological language analysis and the tools at its disposal, it is worth mentioning some new challenges it still faces ahead of it. As these methods of analysis keep entering the internet domain and the large amount of data that resides there, they face new communication and language dynamics that are exclusive to the online network.

### 3. Netspeak

The rise of the internet has impacted language to the extent that new exclusive ways to this environment are being produced. As their emergence and presence became more and more noticeable, these new forms have been called by different names. Crystal (2004) created the term "netspeak", a term that has gained more traction than the others. According to this author, netspeak broadly refers to a new type of language that presents attributes that are unique to the internet. Something that is of special interest is the dual nature that netspeak exhibits. It usually has features from both written and spoken language, creating a kind of hybrid that can be thought of as the product of an expression such as: "Write as you speak".

The author also points out that, when analyzing netspeak, it is easy to see how it has generated variations in different aspects of language, as it has been observed that lexicon and orthography have been the most affected by the emergence of netspeak (Crystal, 2004). Some of this changes are the creation of new words (eg. geek, snob, dork), the compound of words to create different meaning (eg. webcam, clickbait, hotspot), presence of prefixes and suffixes that also create new meanings (eg hypertext, e-book, chatbot, emoticon), abbreviations and acronyms (eg HTML, URL, pls, imo, btw), among others (Hadžiahmetović-Jurida, 2007). On the other hand, the analysis of just orthographic features also reveals interesting aspects of netspeak. For example, Fernández de Molina Ortés (2015) performed an orthographic analysis of the language used in a Spanish social media site and found that there are substitutions, omissions or lengthening of letters in words that,

despite being orthographic errors, reflect an oralized written discourse and a possible economization of the time it takes to create the message, this last part due to the speed of writing a user of social media sites needs in order to communicate effectively in this medium. Similarly, Thangaraj and Maniam (2015) found a wide use of abbreviations when analyzing the netspeak used by students in journal blogs.

Other aspects of language also present changes due to the emergence of netspeak. Wang and Wang (2017) state that, at phonetic level, onomatopoeia are usually found and are used to give a more vivid character to the expression (eg *ZZZZZZZ*); on the other hand, at the syntactic level, there is a marked presence of short sentences with a simple structure, while at the discursive level it is perceived, especially in chat rooms (chatrooms), conversational sequences and a global structure of exchanges. Additionally, since in chatrooms the notion of to whom the message is addressed to can be lost, it is common to place the username of the wanted recipient in front of the message.

If the internet has become a universal medium, then, there is no reason to believe that netspeak has not become universal too. In the era of written language on the internet, in regards to netspeak, it is worth remembering the following:

“It’s simple, useful and easy to interpret graphism is the reason for its universality” (Berlanga and Martínez, 2010, p.51)

Given this universality of netspeak, its inclusion in the psychological analysis of internet texts could prove useful to in improving the comprehension of how psychological phenomena (like personality) are reflected through language on the internet. The next section expands on arguments in favor of this position.

## **4. Netspeak in psychological analysis of texts**

### **4.1. Why include netspeak**

The psychological analysis of internet texts can no longer avoid taking into account netspeak as a relevant factor. It is worth mentioning that not all forms of netspeak are presented uniformly in all virtual spaces. Crystal (2004) comments that there are internet situations where certain forms of netspeak may occur over others, may occur more frequently, or where netspeak may not be present at all. Some examples of these situations, which can also be understood as internet spaces, are discussion forums such as Reddit, social media such as Facebook, Twitter, Instagram, etc., emails, personal blogs, news sites, WhatsApp conversations, among others.

An argument for the dynamic of the presence of netspeak can be found in Biber and Egbert (2016), where the main goal of their research was to identify the patterns linguistic variation among texts from different internet registers. They used a multi-dimensional analysis for lexicogrammatical features on internet texts corpora, organized in several different register and sub-register categories. The authors identified various dimensions of variation and described those patterns of linguistic variation within the registers. Although netspeak was not specifically found or tested, the argument of linguistic variations between different internet register categories impacts netspeak on the basis of whether or not an expression can be correctly classi-

fied as netspeak if it appears in an online environment where linguistic variation is known to happen.

However, some of netspeak’s most distinctive linguistic features, mainly its written-spoken dual style nature, have made it possible for it to be more frequently used on certain online spaces over others, social media being one of these spaces, thus making netspeak expressions more easily identifiable there. A study by Yeo and Ting (2017) analyzed the discourse features of netspeak of Facebook users via their conversations with other users and found features of netspeak use in the conversations that specifically resemble features of speech and writing. For this, the authors argue that, consistent with the original formulations made by Crystal (2004), social media appears to facilitate spoken-like communication while being a primarily written medium. Given these results and their interpretation, the communication dynamics by which social media operates seem to foster an environment where netspeak is more likely to be used.

The inclusion of netspeak expressions and terms in this research are motivated by the meaning that each individual netspeak expression exhibits and how it can be added to a psychological analysis of text. Specifically, including netspeak expressions allows for greater and better amounts of data to be collected from internet sources of text, making it so that these netspeak expressions can be treated as content words that are present within the internet media and, due to this, allows them to have as much value as any other content word for a psychological analysis of any internet text.

These arguments, along with the overall importance that netspeak’s presence in internet language reflects, make up the reasoning behind the methodological decisions that were made and are presented within this research (see Section 5.2.1).

### **4.2. How to include netspeak into psychological analysis**

The approach of psychological analysis centered on dictionaries presents itself as a good opportunity for this task, so does the aforementioned program of psychological analysis of texts, LIWC, and how it allows for the analytic potential of implementing dictionaries focused on netspeak to be tested.

It is important to mention that the 2015 version of the LIWC dictionary in English already takes into account netspeak as a linguistic dimension and also categorizes expressions of this type in their corresponding psychological domains (Pennebaker et al., 2015). However, the dictionary has not been updated since then. In addition, dictionaries validated in other languages do not have a netspeak category or classification of expressions of this type in their psychological dimensions.

This represents two problems. The first arises from the constant change of language on the online network, a change that occurs perhaps faster than usual given the very nature of the internet (Talib Al-Kadi and Ahmed, 2018), making a necessity for dictionaries to be constantly updated. Similarly, this lack of updating affects dictionaries in other languages that have yet to include netspeak.

The second problem comes from the relationship between different languages and the cultures that use them. Al-

though the universality of netspeak as a whole can be inferred given the universality of the internet, netspeak forms do not have this property. There are linguistic expressions whose meanings are closely linked to some cultures and, therefore, may not present a direct semantic equivalent in the languages used by other cultures (Paluszkiwicz-Misiaczek, 2005), so it is reasonable to think that this argument also extends to netspeak and how it is used in different cultures. Therefore, the existence of different netspeak expressions in distinct languages is reaffirmed.

The present study considered as a solution for these problems the expansion of the internal Spanish dictionary with the inclusion of netspeak, which increases the program's analytic potential. Netspeak content was taken from an obtained internet-specific lexicon dictionary for Mexican Spanish (Sánchez et al., 2017), which consists of 247 words used in social network. These terms were incorporated into LIWC's 2007 Spanish dictionary along with a further expansion of netspeak terms that came from analyzing the corpus of tweets and a brainstorm session of ideas (see Section 5.2.1). It is worth nothing that for this study a few reported categories in Yarkoni (2010) were not considered either for the expansion nor for the analysis because the Spanish dictionary did not include them.

The following sections describe the methodology that was carried out aiming to incorporate netspeak into psychological analysis of internet texts. Then, the analysis of personality traits reflection in Twitter users is explained.

## 5. Methodology

Our methodology consists of the following five steps:

- Data gathering through Twitter's API.
- Pre-processing, removing hashtags, mentions, links and emoji, or replacing them with placeholders.
- Elaborations of expanded dictionaries for LIWC in order to deal with netspeak, and running LIWC.
- Data Analysis.

### 5.1. Corpus

Through Twitter's API, we collected a random sample dataset consisting of 121 users, averaging 2000 tweets per user, with a total word count of 2522904 (mean=20850.45, sd=8904.45). This data was collected in September 2019, with the oldest tweets coming from January 2018. The corpus consists entirely of Spanish tweets from Mexico. Because some users wrote about only specific topics like politics, sports or religion, 23 users were not included in the analysis, leaving a sample of 98 users.

### 5.2. Preprocessing

Standard preprocessing methods were used; all words were converted to lowercase, hashtags, mentions, link, retweets and emoji were removed. However, stop words were not eliminated, nor was excessive punctuation.

**Example 1.** b'Por favor ayuden con rt rt rt <https://t.co/u2V53rVFFF>'

**Example 2.** por favor ayuden con rt rt rt

Example 1 shows the text as extracted from Twitter's API, while example 2 is the same tweet after preprocessing.

Furthermore, 23 users were excluded from the analysis as their tweets were entirely composed of weather updates, sports news or other similar, non-personal, content.

#### 5.2.1. The Expanded Dictionary

For our use of LIWC, we obtained the Spanish version of the dictionary (Ramírez-Esparza et al., 2007). Because this version of the Spanish dictionary was created in 2007, it was deemed too outdated in its internet-related content, so the decision was made to incorporate the Spanish Specific Lexicon for Social Networks (Sánchez et al., 2017) onto it. Additionally, a list of words was generated via the extraction of a set of netspeak terms which appeared in the corpus of tweets and via a brainstorm of words intended to complement this set, named Netspeak Lexicon. Both, the Spanish Specific Lexicon for Social Networks and the Netspeak Lexicon were planned to be included into LIWC's dictionary. For this task, we first searched for the presence of the Spanish Specific Lexicon words and the Netspeak Lexicon words within the tweets and applied a key word in context approach to extract their meaning. Words that had a frequency of appearance of less than 10 in the whole corpus of tweets were excluded. The reason was to avoid circumstantial words and therefore, not common enough to be included into the dictionary. The next step was to classify these words into their respective LIWC categories in order for them to be included in the dictionary. To accomplish the categorization process, an inter-annotator procedure with three judges was carried out. First, each one of the judges received a list containing the words that resulted from the previous step, along with each word's definition. The task was to classify those words into LIWC's list of categories. Once this was done, discussions between the three judges were held for the words that had only two judges in agreement or no agreement at all. Only the words whose corresponding category was agreed upon by the three judges during this process were incorporated into LIWC's dictionary, the rest were excluded. From both the Spanish Specific Lexicon Social Networks and the Netspeak Lexicon, a total of 1225 words (being 644 stems) were included into the dictionary. Once the classification process was done, a test-run was made with the expanded LIWC dictionary to make sure that the program recognized and correctly classified the added words.

Results comparing LIWC's performance using the original Spanish 2007 dictionary and the expanded dictionary are presented in Figure 1. These results show that LIWC performs better overall with the expanded dictionary than with the original 2007 Spanish dictionary.

#### 5.2.2. Data Analysis

The chosen personality model was the Big Five personality model that was previously mentioned (Goldberg, 1990).

Due to the sufficiently high amount of samples, we suppose a normal distribution for each of the LIWC features, which then easily allows us to find the 5th and 95th percentile of users in each one of the features.

We contrast these results with the output of Python’s scikit-learn (Pedregosa et al., 2011) implementation of the  $k$ -means algorithm, where the factors are the standardized frequencies of the significant LIWC categories reported on Table 1 of each user. In order to select the  $k$  parameter, we used the silhouette test and the elbow method.

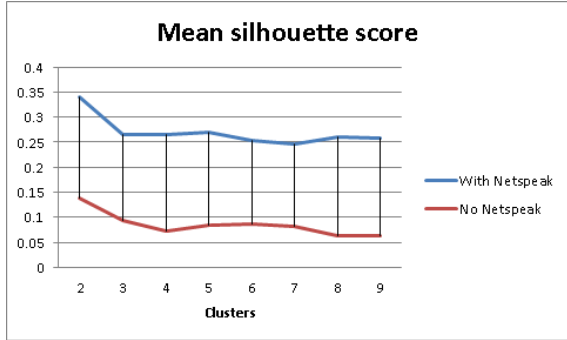


Figure 1: Silhouette test results (Bigger is better)

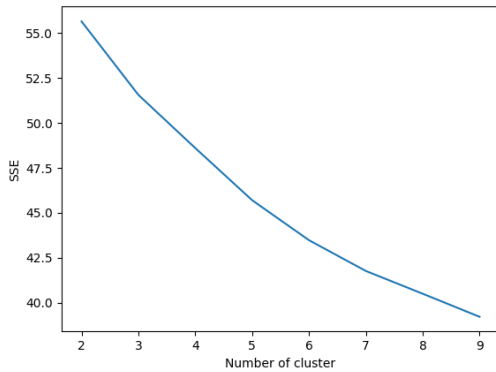


Figure 2: Elbow method results (Looking for inflexion point)

From these tests, we considered values of  $k = 4, 5$  for the analysis, settling on  $k = 5$  for its better representation of the studied phenomenon (see Figure 2).

## 6. Results

Five groups were generated through a clustering process. Table 2 shows the number of users for each cluster, as same as their respective word count mean.

Cluster	n	Word count mean
1	7	14855
2	19	19530
3	32	23792
4	21	21643
5	19	18033

Table 2: Clusters, users count and word count mean

The bar charts illustrate the mean values of the representative LIWC Categories for each possible detected personality trait. The word categories were chosen according to the

significant correlations that Yarkoni (2010) reported. Moreover, the group values are shown together with a global mean of each category, this in order to facilitate the comparison among the different clusters. Figure 3 shows the word categories "You", "We", "Social", "Friends", "Human", "Positive Emotion", "Physical States" and "Sexuality". It can be seen that the first bar for each word category shows the highest values. Therefore, it can be inferred that this cluster contains users who are high in Extraversion. For Conscientiousness it has been primarily reported that this personality trait correlates negatively with almost all word categories, but achievement. Figure 4 illustrates the low values that this cluster has in the mentioned categories. Only the value for Achievement is similar to the global mean ( $M = .58$ ). These results correspond to the previously reported correlations. For this reason, we could assume that this second cluster is formed by users high in Conscientiousness. Figure 5 shows the word categories which have been reported as related with Agreeableness trait. Contrary to expectation, values for "Family" and "Positive emotion" categories are low. Nevertheless, values for "I", "Time", "Motion" and "Home" are clearly high in this cluster, whereas the category "Swear" has a low value. Users in this third cluster might be therefore high in Agreeableness. Openness has been found positively correlated with articles, prepositions. On the other side, it appears to correlate negatively with personal pronouns, negation, social processes, family positive emotion and home words. Contrary to previous findings, Figure 6 shows that values for "Negation", "Social Processes" and "Family" are high. Despite this, "Articles" and "Preposition" have high values, whereas "Pronouns", "Positive Emotion" and "Home" have lower ones in comparison to the other groups. We speculate that this cluster belongs to Openness trait. The representative word categories for Neuroticism are shown in Figure 7. It can be easily seen that this last cluster has high values for "I", "Negation", "Swear", "Negative Emotion", "Anger". The other categories ("Cognitive Processes", "Causation", "Discrepancy") appear to not being very related to this trait as reported. However, the first categories are the main ones for Neuroticism. Consequently, users in this group might be high in this personality trait. Overall it is clear that the five groups are different from each other and that the inclusion of netspeak improves it. Besides that, the comparison of categories values allows to make the inference about which type of personality is represented by each group.

## 7. Discussion

It has been previously documented that language use can give information about people’s psychological characteristics. Different works have reported weak to moderate correlations between personality scores and word categories used in a variety of texts (essay, email, blog, etc.). The current study aimed to test these correlations’ utility for identifying personality traits in a mexican spanish corpus from Twitter.

### 7.1. Netspeak

The Spanish dictionary performance was compared against our expanded version with netspeak. Results confirmed the

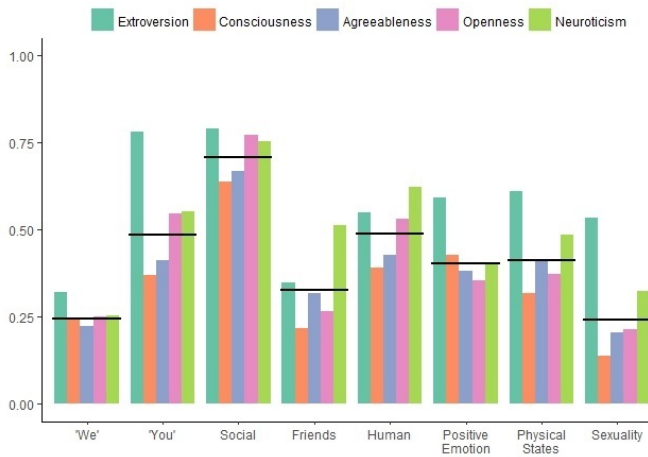


Figure 3: LIWC Categories for Extraversion

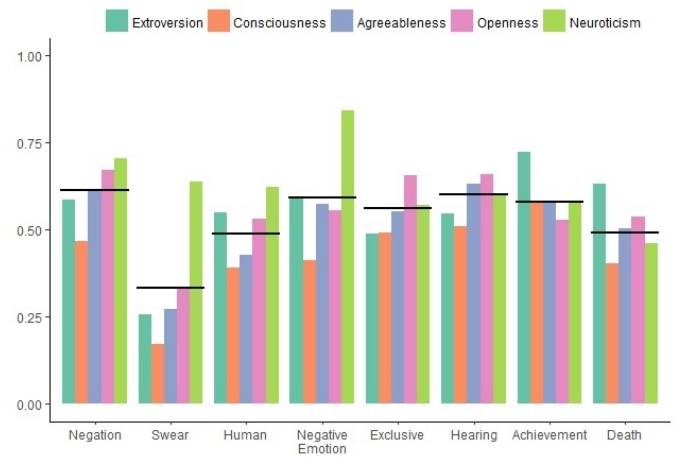


Figure 4: LIWC Categories for Conscientiousness

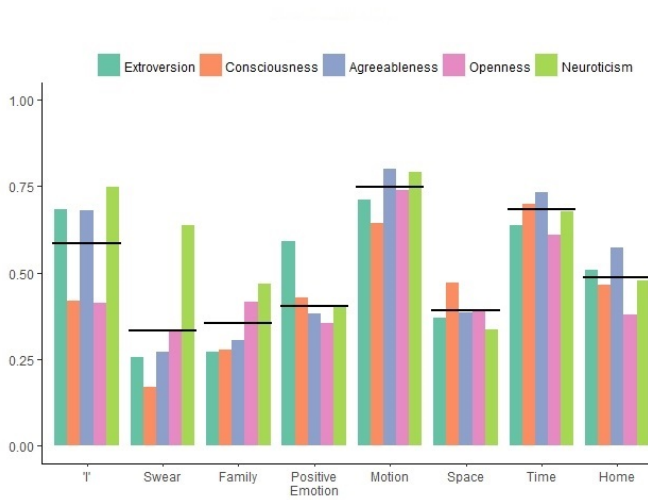


Figure 5: LIWC Categories for Agreeableness

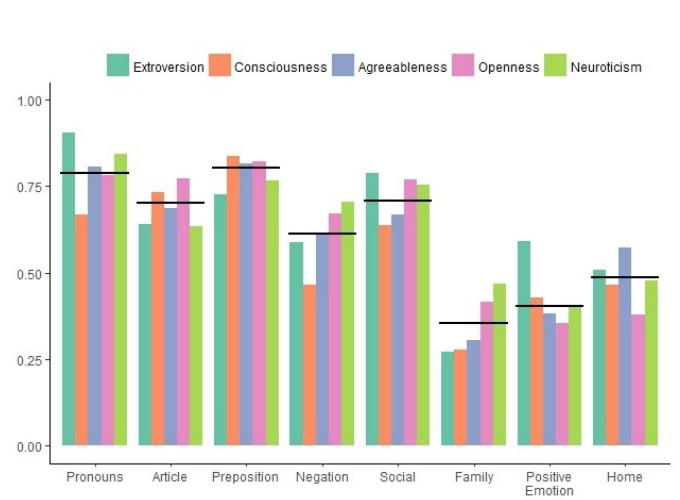


Figure 6: LIWC Categories for Openness

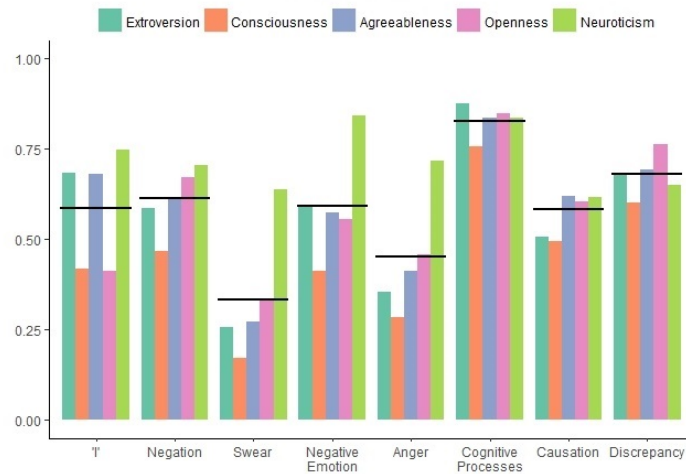


Figure 7: LIWC Categories for Neuroticism

utility of netspeak inclusion, therefore this study highlights the importance of those specific terms used in internet. Social media users have developed new ways for communicating their ideas and thoughts. Netspeak is a language adaptation to the current activities of many young people (Para, 2016). Although many of this kind of terms are not considered as proper language to be in a traditional dictionary, we found evidence that support the necessary inclusion of netspeak in similar studies. The inclusion of these terms must not be just into a separated "Netspeak" category, but into categories whose content and meaning represent. An example is the word "mailob", which comes from the english expression "my love". Because of its meaning, this netspeak term would be clearly included in categories like "Affect", "Positive emotion" and "Social Processes".

## 7.2. Personality expression on Twitter

Clustering procedure classified 98 Twitter users in five groups through an analysis of 34 LIWC categories scores, which were chosen due to its significance reported in Yarkoni (2010). As expected, the first group identified as extraverted shows high values in the representative categories for this trait. Extraversion trait is expressed mainly through the pronouns "we" and "you", this indicating an interest for the other persons. Equally, a high use of words related with social processes, like "Friends" and "Human" categories, indicates an involvement in social interaction. Moreover, in comparison with the other groups, this one appeared to use more words related to positive emotions, physical states and sexuality. Such consistency permits the inference of this group as extraverted. Earlier studies have reported mostly negative correlations between word categories and Consciousness. In this study, a similar pattern was found for this trait. This group has the lowest values in almost all categories, an exception is the "Achievement" category. This ties well with conscientious people, who tend to stay in control, focused and reserved. Agreeableness trait has been described as related with positive emotions and family. However, the values for these categories were not very high. Despite of that, we assumed this third group as representative for this trait because of its high use of first personal pronoun and words related with motion, time and home; categories which have positively correlated with Agreeableness too. It should be noted that this group contained 32 users, being the biggest obtained cluster. Therefore, we argue that this group has more variability compared to the other groups, which might have made difficult to represent the trait as very consistent with previous findings. A trait whose linguistic cues have been not easily captured is Openness (Ireland and Mehl, 2014). Nevertheless, the intellectual aspects of this trait appear to be expressed through the use of articles and prepositions. Such findings were not different in our study. The fourth group had elevated values in these word categories, showing the characteristic formal writing style of open people. Furthermore, low values in "Positive emotion" and "Home" categories are consistent too. Surprising is that this group presents high usage of social words, specifically the ones for family content. This could mean, that family is an important factor for these people's emotionality. Finally, neu-

roticism is one of the better captured traits in words usage. Due to the social rejection of negative emotion expressions, it is broadly accepted that individuals with this trait tend to express more their feelings in Social Media, where they feel more freedom. They show an elevated usage of first personal singular pronoun, as well as words about negative emotions like anger and swearing. Considering the particularities of neuroticism, we conclude that the fifth group includes those users who have this trait.

It should be noted that the inferences drawn from these results are limited. First, we rely on what previous research (Yarkoni, 2010) has found about the relationship between LIWC categories and BigFive personality traits to classify the individuals based on their tweets. Moreover, the present study did not use a self-report personality inventory. Thus, this classification should not be thought of as definitive given this limitation. However, we propose that the groups presented here should be treated as candidates for showing noticeable values in the reported categories which have been found as particularly related to each personality trait. The analysis of writing samples such as tweets is an excellent approach to test what has been found in other studies under constrained circumstances. Even though most reported correlations have been considered as not strong enough, our findings support the linguistic cues as valuable information for personality detection. The assessment of people's personality traits has a variety of useful applications and this study can contribute to further research that aims to develop automatic detection techniques.

## 7.3. Limitations

Although the data obtained in this research can prove itself useful for personality-trait recognition in a Spanish corpus, it is worth mentioning its limitations so that future studies can overcome them and obtain even better data. In this study, we lacked access to personality scores of the analyzed tweets' authors. Therefore, we were not able to prove exactly the accuracy of the method employed. Instead, we relied on correlations that have been consistently reported across studies, specially Yarkoni's (2010) since its huge value derives from its extensive analysis. Nonetheless, the results presented in this study match with a good amount of what the research regarding personality and language has reported. The dictionary-based approach must also be mentioned, because even if the dictionary was expanded with the inclusion of netspeak, new forms of this online language are continuously being created and used, making the constant expansion of the dictionaries a necessity.

## 8. Future Research

The rise of the internet and social media has opened the door for a whole new type of research that focuses on texts to establish relationships between what is written and what it can reveal about the psychology of people. Here, it has been shown that this relationship between text and psychology has been present for a long time but up until recently has gained a new direction that is slowly, but surely, gaining more importance. This new direction takes psychological

analysis of texts to the domain of the internet and its particular uses of language, namely netspeak. Particularly, this work showed the utility of analyzing language use in social media with the aim of detecting personality traits. The potential of language and netspeak as valuable resources for the analysis of multiple dimensions of people's social and personal worlds has been argued for throughout this paper. The presentation of existing software for this task and some methods of analysis being used in an ongoing investigation have also been presented. Incorporating these language elements into the psychological analysis of words through LIWC-style programs represents both a challenge and an opportunity. Moreover, implementation of clustering algorithms for the ongoing investigation remains to be improved. Certainly, there are many other methods and approaches that can be experimented with in order to further develop this emerging field. As more and more research continues to adopt this new direction of psychological internet language analysis, it will become clearer what kind of linguistic aspects are key for understanding the individual and social dynamics reflected in this internet-mediated age.

## 9. Acknowledgment

This work is supported by PAPIIT projects IG400119 and IA401219. We thank for the collaboration of Ricardo Olvera and Silvia Fraustro, from the INEGI, and of Tamara Sierra and Patricia Andrade-Palos from the UNAM.

## 10. Bibliographical References

- Berlanga, I. and Martínez, E. (2010). Ciberlenguaje y principios de retórica clásica. redes sociales: el caso facebook. *Enlace Revista Venezolana de Información, Tecnología y Conocimiento*, 7(2):47–61.
- Biber, D. and Egbert, J. (2016). Register variation on the searchable web: A multi-dimensional analysis. *Journal of English Linguistics*, 44(2):95–137.
- Boyd, R. L. and Pennebaker, J. W. (2015). A way with words: Using language for psychological science in the modern era. In C Dimofte, et al., editors, *Consumer psychology in a social media world*, pages 222–236. Routledge, New York, NY.
- Coppersmith, G., Dredze, M., Harman, C., and Hollingshead, K. (2015). From ADHD to SAD: Analyzing the language of mental health on twitter through self-reported diagnoses. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 1–10.
- Costa, P. and McCrae, R. (1992). Normal personality assessment in clinical practice: The neo personality inventory. *Psychological Assessment*, 4(1):5–13.
- Crystal, D. (2004). *Language and the Internet*. Cambridge University Press.
- Fernández de Molina Ortés, E. (2015). La interacción lingüística en las redes sociales. aspectos ortográficos. In *Estudios sobre ortografía del español*. AXAC.
- Goldberg, L. R. (1981). The search for universals in personality lexicons. In L. Wheeler, editor, *Review of Personality and Social Psychology: Vol. 2*, pages 141–165. Sage, Beverly Hills, CA.
- Goldberg, L. R. (1990). An alternative "description of personality": the big-five factor structure. *Journal of personality and social psychology*, 59 6:1216–29.
- Hadžiahmetović-Jurida, S. (2007). Some distinctive lexical features of netspeak. *Jezikoslovlje*, 8(2):193–210.
- Hancock, J. T., Curry, L. E., Goorha, S., and Woodworth, M. (2008). On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes*, 45(1):1–23.
- Hirsh, J. and Peterson, J. (2009). Personality and language use in self-narratives. *Journal of Research in Personality*, 43(3):524–527.
- Ireland, M. and Mehl, M. (2014). Natural language use as a marker of personality. In T. Holtgraves, editor, *The Oxford Handbook of Language and Social Psychology*, pages 201–2018. Oxford University Press, Oxford.
- Kacewicz, E., Pennebaker, J., Davis, M., Jeon, M., and Graesser, A. (2013). Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology*, 33(2):125–143.
- Majumder, N., Poria, S., Gelbukh, A., and Cambria, E. (2017). Deep learning-based document modeling for personality detection from text. *Intelligent Systems, IEEE*, 32(2):74–79.
- Miller, G. A. (1990). The place of language in a scientific psychology. *Psychological Science*, 1(1):7–14.
- Paluszkiwicz-Misiaczek, M. (2005). Strategies and methods in dealing with culture specific expressions on the basis of Polish-English translations of certain administrative and institutional terms. In *Proceedings from the Eighth Conference of British, American and Canadian Studies*.
- Para, I. (2016). Online communication: Netspeak. the internet as a facilitator for new ways of communication and the impact on our language. In D. Dejica, et al., editors, *Language in the Digital Era: Challenges and Perspectives*, pages 189–200. De Gruyter, Berlin.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Pennebaker, J. W. and Francis, M. E. (1996). Cognitive, emotional, and language processes in disclosure. *Cognition and Emotion*, 10(6):601–626.
- Pennebaker, J. W. and King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296–312.
- Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). *The Development and Psychometric Properties of LIWC2015*. Austin, TX: University of Texas at Austin.
- Qiu, L., Lin, H., Ramsay, J., and Yang, F. (2012). You are what you tweet: Personality expression and perception on twitter. *Journal of Research in Personality*, 46(6):710–718.
- Ramírez-Esparza, N., Pennebaker, J. W., García, F. A., and

- Suriá, R. (2007). Psicología del uso de las palabras: Un programa de computadora que analiza textos en español. *Revista Mexicana de Psicología*, 24(1):85–99.
- Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E. P., and Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE*, 8(9):1–16.
- Talib Al-Kadi, A. M. and Ahmed, R. A. (2018). Evolution of english in the internet age. *Indonesian Journal of Applied Linguistics*, 7(3):727–736.
- Tausczik, Y. and Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.
- Thangaraj, S. and Maniam, M. (2015). The influence of netspeak on students' writing. *Journal of Education and Learning (EduLearn)*, 9:45, 02.
- Wang, D. Z. and Wang, S. S. (2017). Research on features of chatroom netspeak from a stylistic view. *Studies in Literature and Language*, 15(1):43–46.
- Yarkoni, T. (2010). Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, 44(3):363–373.
- Yeo, D. and Ting, S. (2017). Netspeak features in facebook communication of malaysian university students. *Journal of Advanced Research in Social and Behavioural Sciences*, 6(1):81–90.

## 11. Language Resource References

- Fenix Osuna Sánchez and Karen Josette Minjares Ávila and Grigori Sidorov. (2017). *Léxico espectral de redes sociales en español*.

# Using Dependency Syntax-Based Methods for Automatic Detection of Psychiatric Comorbidities

Yannis Haralambous<sup>1</sup>, Christophe Lemey<sup>2</sup>, Philippe Lenca<sup>1</sup>, Romain Billot<sup>1</sup> & Deok-Hee Kim-Dufor<sup>2</sup>

<sup>1</sup> IMT Atlantique, Lab-STICC, UMR CNRS 6285, F-29238, Brest, France

<sup>2</sup> Adolescents and Young Adults Mental Health Department, Brest Medical University Hospital, Brest, France  
& EA 7479 SPURBO, Université de Bretagne Occidentale, Brest, France

## Abstract

This paper presents the early stages of a growing corpus of psychiatric interviews from help seeking patients referred to an early detection and intervention center for psychosis. In order to contribute to the practitioner’s diagnostic, we focus on a new method of automatic comorbidity detection in the corpus. Among the novelties of this method is the fact that it is based on syntactic features of paralinguistic data (interjections and pauses). We use the formalism of dependency syntax, a brief description of which we provide in the paper. Considering the (currently) small size of the corpus, our intention is to prove the applicability of the method rather than to obtain general results about the relevance of syntactic indicators.

## 1. Introduction

According to the 2001 report of the World Health Organization, psychotic disorders (among which schizophrenia) are one of the main public health problems (Anderson, 2019). They are the third disease in terms of disabilities for individuals (Rössler et al., 2005). This chronic and disabling pathology has an important functional and social impact. It may lead to addictive and self-harming behaviors and brings about severe pain in the patients and their relatives. Schizophrenia is a disease that sets in progressively and at various speeds from one individual to another. The symptoms are diverse and unspecific in the stage preceding the prodromal phase (Yung and McGorry, 1996). In addition, these disorders often arise during adolescence which is characterized by upheavals. The evolutive course of schizophrenia is as follows: the premorbid phase, from birth of the patient until the emergence of the first signs; the prodromal phase during which appear scarcely specific first signs of the disease; these unspecific symptoms gradually increase in intensity and specificity during the phase that precedes the clear psychotic symptoms. Eventually, the psychotic phase arises with the known first psychotic signs that determine the onset of psychosis. The active phase of schizophrenia is characterized by a sheaf of very variable symptoms:

1. positive symptoms: delirious ideas and hallucinations;
2. negative symptoms: social withdrawal and cognitive deficits;
3. disorganization syndrome: contact disorder.

About 600,000 people are currently (early 2020) diagnosed with this disease in France<sup>1</sup> and it is notable that one out of two patients attempts to commit suicide during the evolution of the illness (Castelein et al., 2015). Furthermore, marijuana abuse correlates highly with the risk of developing the disease by doubling it (Krebs et al., 2019). It is a complex disease the physiopathology of which remains little known. The current world-wide dominant explanatory

model is the diathesis-stress model that combines two factors: intrinsic vulnerability and stress originating in lived experiences (Howes and McCutcheon, 2017; Bernardo et al., 2017; Pruessner et al., 2017; Millman et al., 2018). Nevertheless, the underlying mechanisms still need to be explored.

The duration between the appearance of the first clear psychotic symptoms and the first access to care is on average two to five years when considered on a world-wide level (differences between various regions being quite important). This period is commonly called “duration of untreated psychosis” (DUP) (Fusar-Poli and others, 2013). Efforts head towards an early treatment and a reduction of the DUP. Indeed, the early identification and rapid interventions during the evolution of a psychotic disorder seem to maximize the therapeutic effects and improve the patients’ quality of life (McGlashan and Johannessen, 1996). During this phase, warning signs prior to the active phase of the disease can be detected, and this results into optimization of care and reduction of the DUP (Olsen and Rosenbaum, 2006). It is these very symptoms that lead patients to medical centers and draw the attention of medical staff for early detection. The patients with unspecific symptoms hinting at the onset of schizophrenia are referred to specialized consultations at centers for early detection of psychosis, for the sake of a further evaluation of each patient’s symptoms. The populations involved are young adults and have previously demonstrated, for the most of them, a suicidal idea or gesture, or behaviors impacting their emotional, social or professional life (Hutton P, 2011). Various studies have resulted into the development of assessment tools (Olsen and Rosenbaum, 2006; Schultze-Lutter, 2009; Yung et al., 2005).

### 1.1. Language Analysis in Psychiatry

Speech, and therefore language, is one of the key elements that clinicians can draw on during psychiatry consultations in order to better understand the patients’ psychological conditions. Psychiatrists are often led to study its phonetic, syntactic and semantic features, which are likely to reveal pathological conditions. Patients with schizophrenia may demonstrate thought disorder, i.e., disorganized

---

<sup>1</sup><https://www.inserm.fr/information-en-sante/dossiers-information/schizophrenie>

thought, which is a characteristic element of this disease. It has been shown that speech analyses can measure thought disorder (Mota et al., 2012). Techniques of computerized speech analyses such as latent semantic analysis, discourse analysis using graph theory and structural discourse analysis have demonstrated a decrease in coherence in patients with schizophrenia correlated with the clinical evaluations and an identical or higher accuracy of diagnosis (Mota et al., 2012; Hoffman et al., 1986; Elvevåg et al., 2007). Through these approaches the first-degree relatives of patients with schizophrenia can be distinguished from control subjects (Elvevåg et al., 2010), and subtly disorganized elements in high-risk patients' speech—which predicts a transition to psychosis—stand out (DeVylder and others, 2014). It has been shown that a combination of semantic and syntactic analyses can predict with reasonable accuracy the transition to schizophrenia and seems to be more efficient than the standard clinical evaluation (SIPS 79%) (Bedi et al., 2015). This method has been replicated in an independent cohort (with an accuracy of 83%) (Corcoran et al., 2018). Prosodic analyses led by different international teams on psychiatric comorbidities have focused mainly on the fundamental frequency (F0) and speech rate (Scherer and Bänziger, 2004; Audibert et al., 2005; van den Broek, 2004; Moore et al., 2003). (Silber-Varod et al., 2016) have in common with our approach the fact that they consider pauses and disfluencies in anxiety comorbidities—nevertheless their work is mainly based on prosodic characteristics, while we focus on syntax.

### 1.2. Psychosis risk assessment

Within the scope of consultations for early detection and intervention, numerous patients are received and have gained access to further assessment of their disorders. The centers for evaluation of risk for psychosis receive patients addressed to them by health, social and care partners who are often helpless before the emergence of a non-constituted psychiatric disorder which manifests itself through an unspecific and polymorphic symptomatology (Le Galudec et al., 2014).

Patients received in our center for early detection (at the Adolescents and Young Adults Mental Health Department, Brest Medical University Hospital) are assessed by a multidisciplinary team including, but not limited to, a psychiatrist, a psychologist, a nurse and a neuropsychologist (Bazziconi et al., 2017). The initial evaluation allows identification of a risk level and the establishment of a personalized care protocol, adapted to the intensity of the disorders. A biannual reevaluation is proposed for two years in order to identify potential aggravations of the disorders and any possible onset of psychosis. This transition of the status of patients “at risk for developing a psychotic disorder” to the onset of a confirmed psychotic pathology is called “transition to psychosis”. Indeed, 24% of the patients at risk develop a psychotic disorder within the following two years and 33% within the following three years (Bazziconi et al., 2017).

It should be noted that the prevalence of comorbidities, especially mood disorder, anxiety and addiction, is very

high (Bazziconi et al., 2017). It is important to identify comorbidities so that patients can have access to appropriate care.

### 1.3. Comorbidities

Comorbidities are evaluated using a standardized clinical interview, i.e., the *Mini International Neuropsychiatric Interview* (MINI) (Sheehan et al., 1998). The following disorders are explored:

- A. Major Depressive Disorder, which we subdivide into:
  - A<sub>1</sub>. Major Depr. Disorder w/o Psychotic disorder
  - A<sub>2</sub>. Major Depr. Disorder w/ Psychotic features
- B. Dysthymia
- C. Suicidality
- D. (Hypo)manic Episode, which we subdivide into:
  - D<sub>1</sub>. Hypomania
  - D<sub>2</sub>. Mania
- E. Panic Disorder
- F. Agoraphobia
- G. Social Phobia
- H. Obsessive Compulsive Disorder
- I. Post-traumatic Stress Disorder
- J. Alcohol Dependence/Abuse
- K. Drug Dependence, which we subdivide into:
  - K<sub>1</sub>. Opioids
  - K<sub>2</sub>. Cocaine
  - K<sub>3</sub>. Cannabis
  - K<sub>4</sub>. Sedatives
- L. Psychotic Disorders
- M. *Anorexia Nervosa*
- N. *Bulimia Nervosa*
- O. Generalized Anxiety Disorder
- P. Antisocial Personality Disorder

We have grouped the psychiatric comorbidities listed above into three groups depending on the nature of the disorders, in order to make it possible to carry out statistical analyses on a limited number of samples: anxiety disorders (**ANX**) (E, F, G, H, I, O); thymic disorders (**THY**) (A, B, C, D); and addictive disorders (**ADD**) (J, K, M, N). Comorbidities L and P have not been explored in the present study.

### 1.4. Project Framework

The results presented in this paper enter into the frame of a research project on informal speech analysis involving all of the patients referred to the early detection and intervention center. The research protocol (NCT03525054) was submitted to, and accepted by, the Institutional Review Board (Comité de Protection des Personnes Est-III, N CPP: 18.04.03). It provides a recording of the initial medical clinical interview and a two-year follow-up.

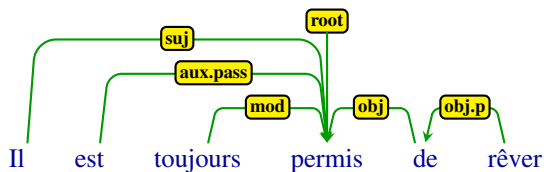
In the following we will first give a short introduction to the specific tool we will be using, namely syntactic dependency relations (§ 2). After that we will describe our corpus and methodology (§ 3) and the results we obtained (§ 4). We conclude with a very short conclusion.

## 2. Dependency Grammars

In the 19th century two linguists from New York, Alonzo Reed and Brainerd Kellogg, introduced a method for rep-

representing syntax relations in a graphical way involving only words occurring in the sentence, and thus avoiding the use of word groups. These diagrams were used in schoolbooks starting from 1877, and lasted way into the 20th century. We don't know whether Lucien Tesnière, a French linguist who studied linguistics in Leipzig, Germany, was aware of their existence, but in the late thirties he started working on a new syntactic theory also based on relations between words, which was published after his death (Tesnière, 1959). At that time, linguists were mainly focused on Chomsky's generative transformational grammars, and so Tesnière's work drew almost no attention outside France. And it would probably stay that way, were there not a researcher from Rand Corporation, David Hays, who introduced Tesnière's ideas to the still young community of computational linguists through a presentation at the notorious UCLA symposium on Machine Translation in 1960 (Hays, 1960), a paper in the *Language* journal in 1964 (Hays, 1964) and, finally, in a book that happened to be the first book dedicated to computational linguistics (Hays, 1967). It was he who introduced the terms *dependency grammar* and *dependency relation*. After Hays, the use of dependency grammars continued to spread and nowadays one can reasonably say that they have largely supplanted methods based on constituents in NLP processes (Kübler et al., 2009; Osborne, 2019). Dependency grammars have already been used in the psychiatric domain, for example in (Tanana et al., 2016) where motivational interviewing sessions have been coded via computer.

In a dependency grammar, each sentence has a *head* (usually the verb) that is the root of a directed tree of *dependency relations*. Edges are directed in such a way that one can draw (directed) paths from each leaf to the root. Every edge has a tag, called *dependency nature*, which describes the relation between the *dependent* (source of the edge) and the *governor* (target of the edge). Here is a dependency tree example, taken from the *French Treebank Corpus* (Abeillé et al., 2003):



We notice in this example (“It is always allowed to dream”) that the participle “permis” is the root of the sentence, and that it governs:

- the pronoun “il” as its subject (suj);
- the verb “est” as its auxiliary verb (aux.pass);
- the adverb “toujours” as a modifier (mod);
- the preposition “de” as its object (obj).

Furthermore, we see that “rêver” is governed by “de” through a prepositional object (obj.p) dependency relation.

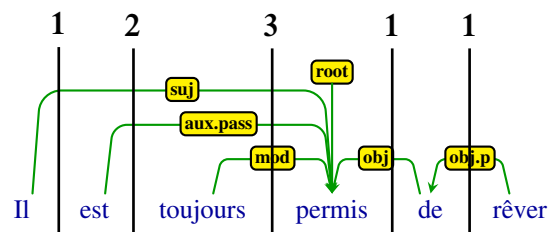
## 2.1. Interstitial Dependency Crossings

Haitao Liu, in (Liu, 2008), explores dependencies from a cognitive point of view and defines a language complexity

measure (*MDD = mean dependency distance*) that quantifies the fact that a sentence such as “The man the boy the woman saw heard left,” although being grammatical, is more difficult to understand than the equivalent “The woman saw the boy that heard the man that left” (the former has an MDD value of 3 and the later an MDD value of 1.4). By the definition of MDD as the average distance between governor and governed, the more “long-distance” dependencies we have, the higher is the MDD value.

Furthermore, dependencies do not overlap, so that we have an irreflexive, asymmetric and transitive relation  $\prec$  between them:  $(a \rightarrow b) \prec (c \rightarrow d)$  when  $(\text{pos}(a) < \text{pos}(c) \text{ and } \text{pos}(d) \geq \text{pos}(b))$  or  $(\text{pos}(a) \leq \text{pos}(c) \text{ and } \text{pos}(d) > \text{pos}(b))$ . In the example above, we have  $(\text{toujours} \rightarrow \text{permis}) \prec (\text{est} \rightarrow \text{permis}) \prec (\text{Il} \rightarrow \text{permis})$ . The relation  $x \prec y$  also implies that  $\text{length}(x) < \text{length}(y)$ .

The binary relation  $\prec$  is a partial order so that we can build a lattice the nodes of which are dependencies and edges represent  $\prec$ . The lengths of paths in this lattice can be visualized in the dependency tree by drawing vertical lines between words:



Here, the fact that  $(\text{toujours} \rightarrow \text{permis}) \prec (\text{est} \rightarrow \text{permis})$ , which is path of length 2 in the lattice, is represented by the fact that the second vertical line crosses two dependencies. Similarly, the path of length 3  $(\text{toujours} \rightarrow \text{permis}) \prec (\text{est} \rightarrow \text{permis}) \prec (\text{Il} \rightarrow \text{permis})$  in the lattice, is represented by the fact that the third vertical line crosses three dependencies. As we see, the number of crossings increases when we approach the root from the left since many dependencies targeting the root accumulate, while on the right, because of adjacency between nodes, the crossing number remains low.

Besides the number of crossings, we also use the nature of crossed dependencies in our calculations, e.g.,  $\{\text{suj}\}$ ,  $\{\text{suj}, \text{aux.pass}\}$ ,  $\{\text{suj}, \text{aux.pass}, \text{mod}\}$  on the left and  $\{\text{obj}\}$  and  $\{\text{obj.p}\}$  on the right, in the example above.

The reason we are interested in interstitial crossings is that in our corpus, besides words we also have *paralinguistic elements*, such as *interjections and pauses*, which occur in interstitial positions.

Our hypothesis is that interstitial positions with a high crossing value are “strategic” and that placing “intruders” (interjections, pauses) in them can be as indicator of some kind of disorder. As we will see, in our small corpus, the combined number and nature of crossed dependencies over a pause or an interjection prove to be comorbidity indicators.

## 2.2. Parsing Informal Text

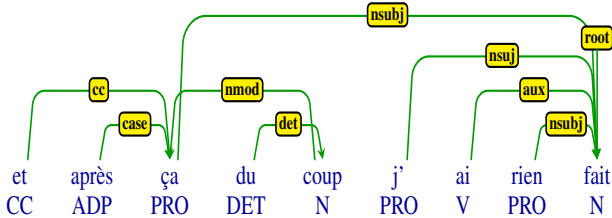
One of the major difficulties of this project was the inability of parsers trained on standard language corpora to parse in-

formal text. We will illustrate this by the syntactic analysis of a typical informal French utterance (by patient #44):

et après ça du coup j'ai rien fait

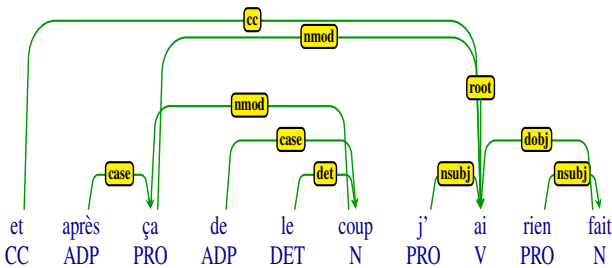
meaning roughly “and after this I haven’t done anything”. This is a complete turn of the patient, located at approx. the end of the first third of the interview.

Here is the result of the (quite popular) spaCy analyzer (Choi et al., 2015):

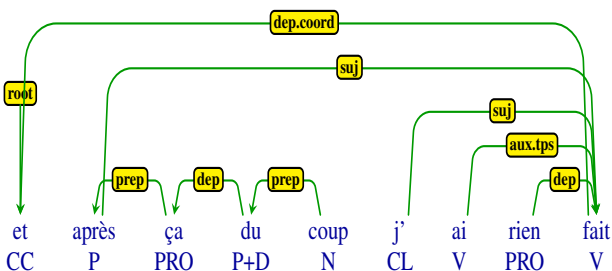


As the reader can see, the word “fait” is taken to be the root of the sentence, but, it carries not a verb tag but a noun tag. Furthermore, the pronoun “rien” is considered as its subject. Also, it is stated that “fait” has a second subject, at the beginning of the sentence, namely the word “ça” (a contraction of “cela”), which is a demonstrative pronoun. It is a strange fact, that “fait” has been chosen to be the root of the sentence but is not tagged as a verb.

The output of the Stanford parser (5/10 2018 version) (Manning et al., 2014) is somehow better:

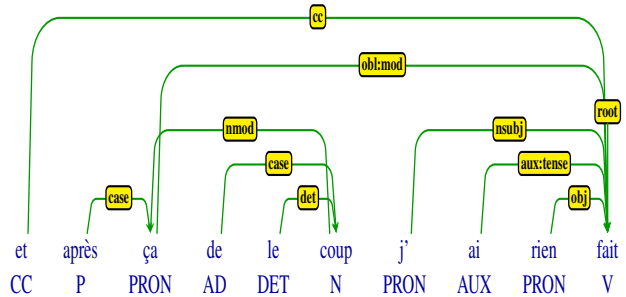


This time it is the auxiliary verb “ai” which is chosen as root of the sentence, and “fait” is its direct object—it is, once again, considered as a noun. Notice that the Stanford parser decomposes “du” into “de le”: a preposition and a determinant. The pronoun “ça” is governed by the root, and the nature of its dependency is that of a modifier, although it would be more natural to take the preposition “après” as modifier and to let it govern “ça” as an prepositional object. The next output comes from the Talismane parser (Urieli, 2013) and is much better in detecting POS tags:



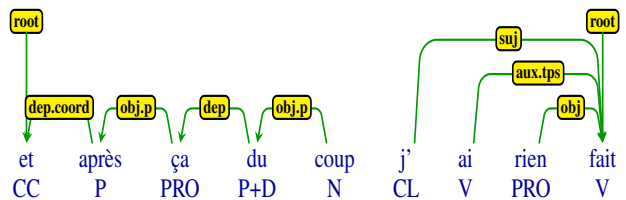
Here it is, surprisingly, the coordination particle “et” which is chosen as root of the utterance, but instead of letting “et” govern the adjacent preposition “après,” Talismane connects “et” with the other extremity of the utterance: the participle “fait” (this time, correctly recognized as a verb). Strangely, the preposition “après” is considered as the subject of “fait”. Otherwise, the parse is correct, and one can notice a split of the utterance into two groups: in the left group, dependencies are adjacent, while in the right group, they are all governed by the verb “fait,” which therefore *acts* like a root, without been *tagged* as a root, probably because Talismane postulates that there can be only one root in an utterance.

The fourth result comes from the very recent package Stanza (with GSD corpus) (Qi et al., 2020). As the reader can see, Stanza avoids all errors made by the other tools we tested: “fait” is detected as being a verb tagged as the root of the sentence, “j” is its subject and “rien” its object:



But again, similarly to the other tools, Stanza considers that there must be only a single root in a sentence and a single dependency tree, therefore we have long (and not very relevant) dependencies between “et” and “fait” and “ça” and “fait”.

The fifth and last result we present is the output of grew (Guillaume and Perrier, 2015), a tool based on the technique of graph rewriting (Bonfante et al., 2018):



Here, the POS tags are the same as in the Talismane example, since the syntax parser grew uses Talismane as its preliminary POS tagger. We notice immediately that grew has indeed recognized the two groups (which Talismane has also noticed but was unable to separate) and has tagged them separately, each one with its own root. Besides the fact that “du coup” could be considered as a secondary interjection and governed by “et,” this analysis is by far the most pertinent, and for this reason we have chosen this tool for our project.

Before closing this section we would like to insist on the fact that this comparison of five renowned syntax parsers is de facto unfair since we are using them for something for which they have not been developed, namely for the

Table 1: Comorbidity values of our corpus’ patients

ID	Gender	Duration	A <sub>1</sub>	A <sub>2</sub>	B	C	D <sub>1</sub>	D <sub>2</sub>	E	F	G	H	I	J	K <sub>1</sub>	K <sub>2</sub>	K <sub>3</sub>	K <sub>4</sub>	L	M	N	O	P	THY	ANX	ADD	
15	F	47’29’’	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	M	47’45’’	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	3	0
23	F	43’50’’	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	3	1	0
25	M	30’59’’	1	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	3	1	0
27	M	25’05’’	1	0	0	2	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	3	2	1
28	M	27’26’’	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	2	2	0
30	M	63’08’’	0	1	0	1	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	2	1	1
44	M	43’13’’	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	1	0	0	0	0	0	1	1	2

analysis of transcribed *informal speech utterances*. Possibly some of them would give better results if the utterance had been properly punctuated, but we preferred—as does also (Blanche-Benveniste, 1990)—not to use punctuation since it is not introduced by the patient but by the secretary transcribing the interview.

### 3. Corpus and Methodology

#### 3.1. The corpus

Our corpus consists of eight patient interviews, of a duration between 25 and 63 minutes. In Table 1 the reader can see characteristics of the patients (gender) and of the interview (duration) as well as the values of standard comorbidities (A-P). As mentioned above, we have grouped comorbidities in three groups, as follows:

**THY** *Thymic Disorders* (comorbidities A, B, C, D);

**ANX** *Anxiety Disorders* (comorbidities E, F, G, H, I, O);

**ADD** *Addictive Disorders* (comorbidities J, K, M, N).

Interviews have been recorded and transcribed by a medical secretary, following conventions for interjections and paralinguistic respiration given in (Bigi, 2015). These transcripts have been proofread and edited by an independent proofreader.

#### 3.2. Methodology

Each recorded interview is segmented into turns between caregiver and patient, and only the latter is kept for further analysis.

Once the transcription has been carefully verified, the two data streams (sound and text) are supplied to SPPAS (Bigi, 2015), which produces a file with timestamped phonemes, words in standard orthography and words in phonemic representation.

But SPPAS does not capture pauses and gaps/lapses. We therefore use Praat (Boersma and Weenink, 2001) on a noise-filtered version of the sound file to detect pauses and gaps/lapses, and then introduce them into the timestamped phoneme and word data. Praat provides us also with energy, pitch, F1 and F2 data, which align with phonemes and words.

In a different data flow (see Fig. 1) we remove interjections from the transcribed text and perform POS tagging on the result using Talismane (Urieli, 2013; Urieli and Tanguy, 2013), followed by dependency parsing using *grew* (Guillaume and Perrier, 2015). This process provides us with (relatively clean) CoNLL data.

We then align the two data flows (data provided from SPPAS and data in CoNLL form) using the Needleman-Wunsch algorithm as implemented in *bioPython*. This provides us with a timestamped version of the CoNLL data. We then use the timestamps of pauses and interjections to study their crossing with syntactic dependencies.

#### 3.3. Definition and Rationale of PIDC and IIDC

Let us consider the dependency syntax forest<sup>2</sup> of a given utterance  $P$ . As the reader can see in Fig. 2, in an utterance (taken from patient #44) such as

les mamans s’inquiètent toujours et les mamans  
embêtent toujours ce genre-là quoi soyez zen  
écoutez

words “quoi” and “écoutez” are not connected to the syntax tree of the two coordinated sentences “les mamans s’inquiètent toujours” and “les mamans embêtent toujours ce genre-là” (“quoi” and potentially “ce genre-là” could also be considered as secondary interjections). We therefore do not have a single syntax tree but tree fragments of varying sizes.

The primary interjection “hein” has not been used for the calculation of syntax dependencies, since we have removed it earlier in the process and reintroduced it afterwards. Indeed, we remove all primary interjections in order to obtain dependencies that are closer to the speaker’s intention (and to avoid misinterpretation by the syntax parser which has been trained on a corpus without interjections).

The *IIDC* (Interjection Interstitial Dependency Crossings) method consists in re-introducing interjections into the syntax tree by using their timestamps and observing crossings with dependency relations. As the reader can see in Fig. 2, primary interjection “hein” crosses a dependency relation between the noun “mamans” acting as a subject, and the verb “inquiètent,” which is the root of the tree fragment. Another interjection (secondary, this time), “quoi,” is not crossing any dependency relation since it is located between distinct syntax trees in the forest.

We act similarly for pauses: *PIDC* (Pause Interstitial Dependency Crossings) is the same method applied to pauses (i.e., silences internal to each patient’s turn): by their timestamps we align them with the syntax tree fragments and find crossings between them and dependency relations.

Our hypothesis is the following:

<sup>2</sup>We call it a *forest* because of the lack of connectivity, as in the example in §2.2.

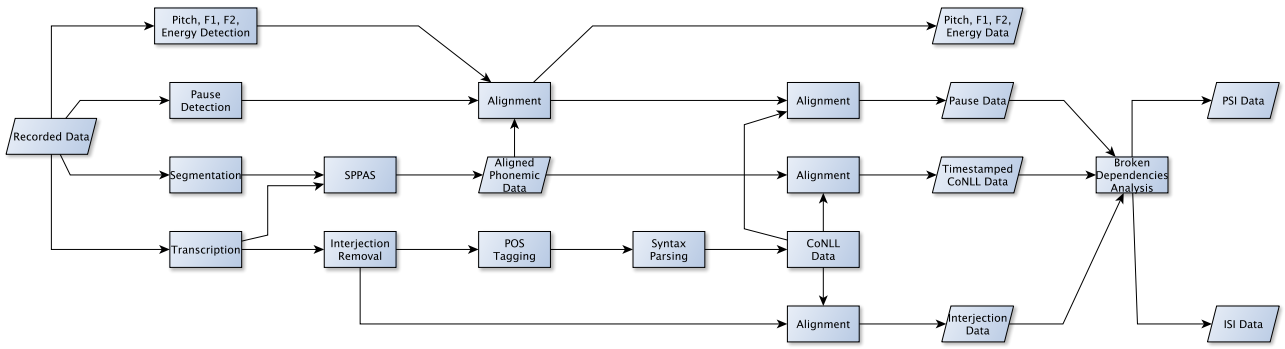


Figure 1: The process of data extraction

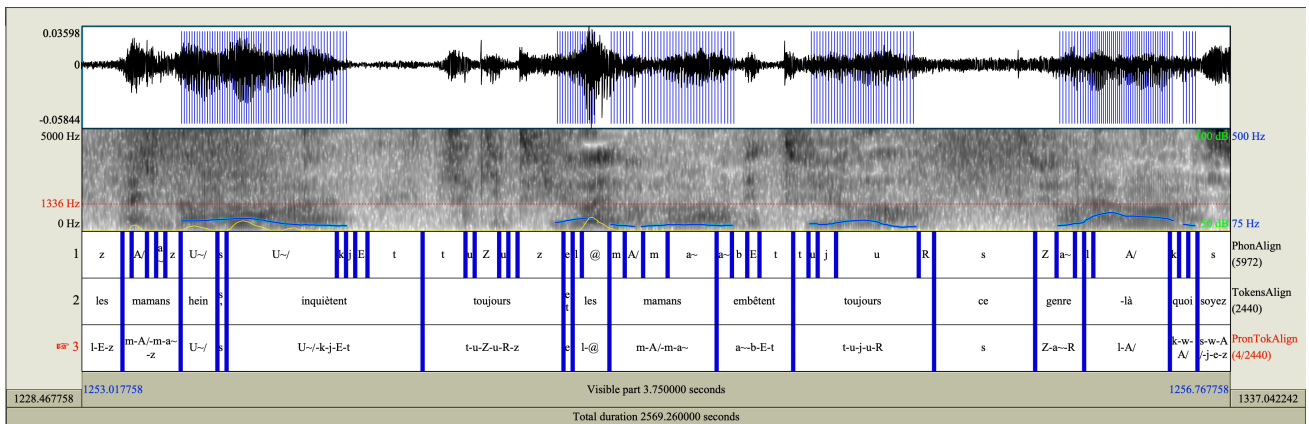


Figure 2: The same utterance (patient #44) visualized in Praat (phonemic alignment) and annotated by syntactic dependencies. The primary interjection “hein” crosses the “suj” dependency “mamans” → “inquiètent”. The secondary interjection “quoi” does not cross any dependency relation.

Interjection Interstitial Dependency Crossings and Pause Interstitial Dependency Crossings can serve as indicators of the patient’s linguistic disorganization.

We measure:

1. the number of dependencies crossing interjections or pauses;
2. the nature of these dependency relations.

We define

$$PIDC := (\#crossings) \times \frac{\text{pause duration}}{\text{utterance duration}};$$

$$IIDC := (\#crossings) \times \frac{\text{interjection duration}}{\text{utterance duration}},$$

and for a given set of dependency relations  $S$  we define:

$$PIDC_S := (\#crossings \text{ in } S) \times \frac{\text{pause duration}}{\text{utterance duration}};$$

$$IIDC_S := (\#crossings \text{ in } S) \times \frac{\text{interjection duration}}{\text{utterance duration}}.$$

Besides PIDC and IIDC, we have calculated  $PIDC_S$  and  $IIDC_S$  for three sets of dependency relations:  $\{\text{det}, \text{suj}\}$ ,  $\{\text{det}\}$ ,  $\{\text{obj.p}\}$  and  $\{\text{suj}\}$ . The justification of these choices is as follows.

In the French Treebank Corpus (Abeillé et al., 2003) (which is the most important publicly available dependency-annotated French corpus), among the most frequent relations we note the following:

nature	frequency	avg dist. dep./gov.
mod	120,741	4.1937
obj.p	90,400	1.7511
det	85,154	1.1987
suj	35,402	4.2315

The “mod” (modifier) dependency is very frequent but can take various forms: in 26% of cases the dependent word is an adjective, in 22% of cases a preposition, in 20% of cases an adverb and in 18% of cases a word, and all of these can be located at a certain distance from their governor, therefore the existence of a pause or an interjection between dependent and governor is not necessarily significant.

On the contrary, the “obj.p” (prepositional object) dependency is actually the equivalent of *case government* (for cased languages) and therefore, according to (Osborne, 2019, p. 142), it is technically a type of morphological dependency rather than a syntactic one. It is very stable in terms of POS tag (86% of its dependents are nouns) and the distance between dependent and governor is quite small (1.7511 in average). Its morphological nature and its positional characteristics lead us to formulate the hypothesis that the crossing of an interjection or of a pause with an obj.p dependency is very likely to reveal disorganization. The “det” (determinant) dependency is also quite suitable to reveal disorganization: the list of determinants is very small and they are very close to their governor (1.1987 in average, that is the smallest average distance above all relations). Finally the “subj” relation is an important one since (expect in the imperative mode) every French verb necessarily has a subject. We investigate this dependency relation, despite its high distance between dependent and governor (4.2315 in average).

#### 4. Results

We performed a Spearman correlation test on the comorbidity values of the three categories **THY**, **ANX** and **ADD** vs. the various indicators we calculated. Here are the most pertinent results.

We display below the Spearman rho value (and p-value to attest the significance of the results) for comorbidity groups and crossing between pauses/interjections and specific dependency groups:

pause/interj.	{depend.}	group.	rho	p-value
pause	{obj.p}	<b>ADD</b>	<b>0.8660</b>	0.0054
pause	{det}	<b>ADD</b>	0.7735	0.0254
pause	{det,suj}	<b>ADD</b>	0.5770	0.1340
pause	{subj}	<b>ANX</b>	-0.5086	0.1980
pause	all	<b>THY</b>	0.7042	0.0512
interjection	{obj.p}	<b>ADD</b>	0.8248	0.0117
interjection	{det}	<b>ADD</b>	0.7285	0.04
interjection	{det,suj}	<b>ANX</b>	-0.8247	0.0117
interjection	{subj}	<b>ANX</b>	<b>-0.8450</b>	0.0080
interjection	all	<b>THY</b>	-0.6730	0.0671

We notice that for {obj.p} and {det} we get similar behavior for pauses and interjections, even though these two paralinguistic phenomena are quite distinct and have been measured in different ways (pauses have been measured globally by Praat, while interjections have been included by the secretary in the transcription, removed afterwards in order to perform syntax analysis, and re-introduced by their timestamps in SPPAS).

Also we notice that pauses or interjections crossing the {obj.p} dependency are a very strong indicator ( $\rho > 0.82$ ) of the **ADD** group, with a high significance ( $p = 0.012$ ). The {det} dependency also has a consistent behavior (rho around 0.75 with a p-value between 0.025 and 0.04) and, again, targets the **ADD** group.

For the other dependencies, values reveal different behaviors: while pauses crossing {det,suj} or {subj} give insignificant results (p-value  $> 0.13$ ), interjections combined with

{det,suj} and {subj} give very high results, but target negatively the **ANX** group ( $\rho < 0.824$  with p-value = 0.012). These results can be expressed as follows:

Members of the **ADD** group tend to place pauses or interjections between preposition and governed noun or between determinant and noun governing it.

Members of the **ANX** group tend to place interjections (but not pauses) between determinant and noun governing it, or between subject and verb governing it.

The first result may reflect the high prevalence of addictive behaviors in patients at risk for psychosis (Valmaggia et al., 2014). As presented previously, the crossing of an interjection or of a pause between preposition and noun or between determinant and noun is very likely to reveal disorganization which is one of the psychotic symptoms often found in at-risk patients (Fusar-Poli and others, 2013). Moreover, the intensity of these psychotic symptoms is correlated with the importance of addictive behaviors (Korver et al., 2010). The second result can be explained by a tendency in anxious patients to avoid leaving gaps, particularly in the context of a conversation where the individual is subject to the judgment of his interlocutor, exactly as would stuttering patients (Iverach and Rapee, 2014).

#### 5. Conclusion

These results show that it is possible to use natural language processing to explore psychiatric comorbidities using linguistic markers. The dependencies and their crossing with pauses and interjection seem to be of particular interest to study in this field. We intend to continue the exploration of linguistic markers following different modalities (semantic, syntactic, prosodic) in order to identify relevant markers for clinical practice.

#### 6. Bibliographical References

- Abeillé, A., Clément, L., and Toussnel, F. (2003). Building a treebank for French. In *Treebanks*, pages 165–187. Kluwer.
- American Psychiatric Association APA. (2013). DSM–5: Diagnostic and Statistical Manual of mental disorders, 5th edition. *American Psychiatric Publishing*.
- Anderson, K. (2019). Towards a public health approach to psychotic disorders. *Lancet Public Health*, 4(5):e212-3.
- Audibert, N., Aubergé, V., and Rilliard, A. (2005). The prosodic dimensions of emotion in speech: the relative weights of parameters. In *European Conference on Speech Communication and Technology*, pages 525–528. ISCA.
- Bazziconi, P., Lemey, C., Bleton, L., and Walter, M. (2017). CEVUP program: An analytical epidemiological cohort study. *European Psychiatry*, 41:S729.
- Bedi, G., Carrillo, F., Cecchi, G., Fernández-Slezak, D., Sigman, M., Mota, N., Ribeiro, S., Javitt, D., Copelli, M., and Corcoran, C. (2015). Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophrenia*, 1:15030.

- Bernardo, M., Bioque, M., Cabrera, B., Lobo, A., González-Pinto, A., Pina, L., Corripio, I., Sanjuán, J., Mané, A., Castro-Fornieles, J., Vieta, E., Arango, C., Mezquida, G., Gassó, P., Parellada, M., Saiz-Ruiz, J., Cuesta, M. J., Mas, S., and PEPs GROUP. (2017). Modelling gene–environment interaction in first episodes of psychosis. *Schizophr. Res.*, 189:181–189.
- Bigi, B. (2015). SPPAS – Multi-lingual Approaches to the Automatic Annotation of Speech. *The Phonetician – International Society of Phonetic Sciences*, 111–112:54–69.
- Blanche-Benveniste, C. (1990). *Le français parlé. Études grammaticales*. Éditions du CNRS, Paris.
- Boersma, P. and Weenink, D. (2001). PRAAT, a system for doing phonetics by computer. *Glott International*, 5(9–10):341–347.
- Bonfante, G., Guillaume, B., and Perrier, G. (2018). *Application of Graph Rewriting to Natural Language Processing*, volume 1 of *Logic, Linguistics and Computer Science Set*. ISTE Wiley.
- Bowie, C., Gupta, M., and Holshausen, K. (2013). Cognitive remediation therapy for mood disorders: Rationale, early evidence, and future directions. *Can. J. Psychiatry*, 58/6:319–325.
- Castelein, S., Liemburg, E., de Lange, J., van Es, F., Visser, E., Aleman, A., Bruggeman, R., and Knegtering, H. (2015). Suicide in Recent Onset Psychosis Revisited: Significant Reduction of Suicide Rate over the Last Two Decades — A Replication Study of a Dutch Incidence Cohort. *PLoS One*, 10:e0129263.
- Choi, J. D., Tetreault, J., and Stent, A. (2015). It depends: Dependency parser comparison using a Web-based evaluation tool. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 387–396. Association for Computational Linguistics.
- Corcoran, C., Carrillo, F., Fernández-Slezak, D., Bedi, G., Klim, C., Javitt, D., Bearden, C., and Cecchi, G. (2018). Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry*, 17(1):67–75.
- DeVylder, J. E. et al. (2014). Symptom trajectories and psychosis onset in a clinical high-risk cohort: the relevance of subthreshold thought disorder. *Schizophr Res*, 159:278–283.
- Elvevåg, B., Foltz, P. W., Weinberger, D. R., and Goldberg, T. E. (2007). Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia. *Schizophr. Res*, 93:304–316.
- Elvevåg, B., Foltz, P. W., Rosenstein, M., and DeLisi, L. E. (2010). An automated method to analyze language use in patients with schizophrenia and their first-degree relatives. *J. Neurolinguistics*, 23:270–284.
- Fusar-Poli, P. et al. (2013). The psychosis high-risk state: a comprehensive state-of-the-art review. *JAMA Psychiatry*, 70:107–120.
- Guillaume, B. and Perrier, G. (2015). Dependency parsing with graph rewriting. In *International Conference on Parsing Technologies*, pages 30–39.
- Hays, D. (1960). Grouping and dependency theories. In *Proceedings of the National Symposium on Machine Translation*, pages 257–266. UCLA.
- Hays, D. (1964). Dependency theory: A formalism and some observations. *Language*, 40:159–525.
- Hays, D. (1967). *Introduction to computational linguistics*. Macdonald & co.
- Hoffman, R. E., Stopek, S., and Andreasen, N. C. (1986). A comparative study of manic vs schizophrenic speech disorganization. *Arch. Gen. Psychiatry*, 43:831–838.
- Howes, O. D. and McCutcheon, R. (2017). Inflammation and the neural diathesis-stress hypothesis of schizophrenia: A reconceptualization. *Transl. Psychiatry*, 7:1024.
- Hutton P, Bowe S, P. S. F. S. (2011). Prevalence of suicide risk factors in people at ultra-high risk of developing psychosis: a service audit. *Early Interv Psychiatry*, 5, nov.
- Iverach, L. and Rapee, R. M. (2014). Social anxiety disorder and stuttering: current status and future directions. *J. Fluency Disord.*, 40:69–82.
- Knight, M. J. and Baune, B. T. (2018). Cognitive dysfunction in major depressive disorder. *Curr. Opin. Psychiatry*, 31:26–31.
- Korver, N., Nieman, D. H., Becker, H. E., van de Fliert, J. R., Dingemans, P. H., de Haan, L., Spiering, M., Schmitz, N., and Linszen, D. H. (2010). Symptomatology and neuropsychological functioning in cannabis using subjects at ultra-high risk for developing psychosis and healthy controls. *Australian & New Zealand Journal of Psychiatry*, 44(3):230–236.
- Krebs, M.-O., Kebir, O., and Jay, T. (2019). Exposure to cannabinoids can lead to persistent cognitive and psychiatric disorders. *European Journal of Pain*, 23.
- Kübler, S., McDonald, R., and Nivre, J. (2009). *Dependency Parsing*, volume 2 of *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool.
- Le Galudec, M., Cornily, G., Garlantézec, R., Stéphan, F., Alavi, Z., and Walter, M. (2014). Evaluation of GPs diagnostic knowledge and treatment practice in detection and treatment of early schizophrenia: a French postal survey in brittany. *Social Psychiatry & Psychiatric Epidemiology*, 49:69–77.
- Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9:159–191.
- Malhi, G. S., Byrow, Y., Fritz, K., Das, P., Baune, B. T., Porter, R. J., and Outhred, T. (2015). Mood disorders: neurocognitive models. *Bipolar Disorders*, 17(S2):3–20.
- Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.
- McGlashan, T. H. and Johannessen, J. O. (1996.). Early detection and intervention with schizophrenia: rationale. *Schizophr. Bull.*, 22:201–222.
- Millman, Z. B., Pitts, S. C., Thompson, E., Kline, E. R., Demro, C., Weintraub, M. J., DeVylder, J. E., Mittal, V. A., Reeves, G. M., and Schiffman, J. (2018). Perceived social stress and symptom severity among help-

- seeking adolescents with versus without clinical high-risk for psychosis. *Schizophr. Res.*, 192:364–370.
- Modinos, G. and McGuire, P. (2015). The prodromal phase of psychosis. *Current Opinion in Neurobiology*, 30:100–105.
- Moore, E., Clements, M., Peifer, J., and Weisser, L. (2003). Analysis of prosodic variation in speech for clinical depression. In *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 3, pages 2925–2928.
- Mota, N. B., Vasconcelos, N. A. P., Lemos, N., Pieretti, A. C., Kinouchi, O., Cecchi, G. A., Copelli, M., and Ribeiro, S. (2012). Speech graphs provide a quantitative measure of thought disorder in psychosis. *PLOS ONE*, 7(4):e34928.
- Olsen, K. A. and Rosenbaum, B. (2006). Prospective investigations of the prodromal state of schizophrenia: assessment instruments. *Acta Psychiatr. Scand.*, 113:273–282.
- Osborne, T. (2019). *A Dependency Grammar of English*. John Benjamins, Amsterdam/Philadelphia.
- Pruessner, M., Cullen, A. E., Aas, M., and Walker, E. F. (2017). The neural diathesis–stress model of schizophrenia revisited: An update on recent findings considering illness stage and neurobiological and methodological complexities. *Neurosci. Biobehav. Rev.*, 73:191–218.
- Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Association for Computational Linguistics (ACL) System Demonstrations*. arXiv:2003.07082.
- Rössler, W., Joachim Salize, H., van Os, J., and Riecher-Rössler, A. (2005). Size of burden of schizophrenia and psychotic disorders. *European Neuropsychopharmacology*, 15:399–409.
- Scherer, K. R. and Bänziger, T. (2004). Emotional expression in prosody: a review and an agenda for future research. In *The Speech Prosody Conference*.
- Schultze-Lutter, F., Ruhrmann, S., Berning, J., Maier, W., and Klosterkötter, J. (2010). Basic symptoms and Ultra-High Risk criteria: Symptom development in the initial prodromal state. *Schizophr Bull.*, 36/1:182–191.
- Schultze-Lutter, F. (2009). Subjective symptoms of schizophrenia in research and the clinic: The basic symptom concept. *Schizophr. Bull.*, 35:5–8.
- Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., Hergueta, T., Baker, R., and Dunbar, G. C. (1998). The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *J. Clin. Psychiatry*, 59 Suppl 20:22–33.
- Silber-Varod, V., Kreiner, H., Lovett, R., Levi-Belz, Y., and Amir, N. (2016). Do social anxiety individuals hesitate more? The prosodic profile of hesitation disfluencies in Social Anxiety Disorder individuals. In *Proceedings of the Speech Prosody Conference*, pages 1211–1215.
- Tanana, M., Hallgren, K. A., Imel, Z. E., Atkins, D. C., and Srikumar, V. (2016). A comparison of natural language processing methods for automated coding of motivational interviewing. *J Subst Abuse Treat.*, 65:43–50, June. doi:10.1016/j.jsat.2016.01.006.
- Tesnière, L. (1959). *Éléments de syntaxe structurale*. Klincksieck, Paris.
- Urieli, A. and Tanguy, L. (2013). L’apport du faisceau dans l’analyse syntaxique en dépendances par transitions: études de cas avec l’analyseur Talismane. In *Actes de la 20<sup>e</sup> conférence sur le Traitement Automatique des Langues Naturelles (TALN’2013)*, pages 188–201.
- Urieli, A. (2013). *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Ph.D. thesis, Université de Toulouse II le Mirail.
- Valmaggia, L. R., Day, F. L., Jones, C., Bissoli, S., Pugh, C., Hall, D., Bhattacharyya, S., Howes, O., Stone, J., Fusar-Poli, P., and et al. (2014). Cannabis use and transition to psychosis in people at ultra-high risk. *Psychological Medicine*, 44(12):2503–2512.
- van den Broek, E. L. (2004). Emotional prosody measurement (EPM): a voice-based evaluation method for psychological therapy effectiveness. *Stud. Health Technol. Inform.*, 103:118–125.
- Yung, A. and McGorry, P. (1996). The prodromal phase of first-episode psychosis: Past and current conceptualizations. *Schizophrenia Bulletin*, 22(2):353–370.
- Yung, A. R., Yung, A. R., Yuen, H. P., McGorry, P. D., Phillips, L. J., Kelly, D., Dell’olio, M., Francey, S. M., Cosgrave, E. M., Killackey, E., Stanford, C., Godfrey, K., and Buckby, J. (2005). Mapping the onset of psychosis: The comprehensive assessment of at-risk mental states. *Australian & New Zealand Journal of Psychiatry*, 39(11-12):964–971.



## Author Index

- Agurto, Carla, 1  
Albert, Pierre, 67  
Alexandersson, Jan, 46  
Alloatti, Francesca, 16  
Álvarez-Marquina, Agustín, 74  
Ash, Sharon, 60
- Bel-Enguix, Gemma, 132  
Billot, Romain, 142  
Bolioli, Andrea, 16  
Bosca, Alessio, 16  
Bustos, Alfredo, 132
- Cabrera-Mora, Adriana, 132  
Canfield, Allison, 54  
Cecchi, Guillermo, 1  
Chandler, Chelsea, 9  
Cheng, Jian, 9  
Cho, Sunghye, 60  
Cohen, Alex S., 9  
Cognon, Louise-Amélie, 108  
Cucchiaroni, Catia, 78  
Cuteri, Vittoria, 34
- Dunfield, Katherine Ann, 24
- Elvevåg, Brita, 9  
Eyigoz, Elif, 1
- Fairon, Cédric, 108  
Foltz, Peter, 9  
Francia, Valentina, 34  
de la Fuente, Sofía, 67
- Holmlund, Terje B., 9  
Hubers, Ferdy, 78
- Ivtushok, Elizaveta, 116
- König, Alexander, 46  
Khudyakova, Mariya, 98, 116  
Kim-Dufor, Deok-Hee, 142
- Lemey, Christophe, 142  
Lenca, Philippe, 142
- Liberman, Mark, 60  
Lindsay, Hali, 46  
Luz, Saturnino, 67
- Malaspina, Elisabetta, 34  
Merkus, Julia, 78  
Minori, Giulia, 34  
Moscano, Filomena, 34
- Neumann, Günter, 24, 88  
Nevler, Naomi, 60  
Nöth, Elmar, 1
- Orozco, Juan Rafael, 1  
Osornio-Arteaga, Alejandro, 132
- Palacios-Alonso, Daniel, 74  
Parmeggiani, Antonia, 34  
Pietrowicz, Mary, 1  
Prud'hommeaux, Emily, 54
- Ravelli, Andrea Amelio, 124  
Romo-Anaya, Ana-Miriam, 132  
Rossi, Francesca, 34  
Rusz, Jan, 1  
Ryazanskaya, Galina, 98
- Seminck, Olga, 108  
Shellikeri, Sanjana, 60  
Shulgina, Kira, 116  
Sierra, Gerardo, 132  
Silva-Cuevas, Víctor, 132  
Silverman, Laura B., 54  
Skodda, Sabine, 1  
Strik, Helmer, 78
- Tamburini, Fabio, 34  
Toldova, Svetlana, 116  
Tröger, Johannes, 46  
Tsanas, Athanasios, 74
- Vechkaeva, Anna, 88
- Yang, Christine, 54