DEPARTMENT OF PHILOSOPHY,
LINGUISTICS AND THEORY OF SCIENCE

DETERMINING LINGUISTIC PREDICTORS FOR THE CLASSIFICATION OF SUBJECTIVE COGNITIVE IMPAIRMENT AND MILD COGNITIVE IMPAIRMENT USING MACHINE LEARNING

Tian Wang

Master's Thesis:	30 credits
Programme:	Master's Programme in Language Technology
Level:	Advanced level
Semester and year:	Spring, 2019
Supervisor:	Charalambos (Haris) Themistocleous
Examiner:	Richard Johansson
Report number:	(number will be provided by the administrators)
Keywords:	Mild cognitive impairment, Subjective cognitive impairment, natural language processing, support vector machine, neural networks

Abstract

Introduction

Mild Cognitive Impairment (MCI) is a neurological condition characterized by cognitive decline greater than expected for an individual's age and education level. Subjective Cognitive Impairment (SCI) is a self-reported decline in cognitive abilities but not clinically identified as MCI. Individuals with MCI remain functional in their daily activities (Petersen et al., 1999) and are characterized by different deterioration rates depending on the evaluation methods employed. More than 50% of these individuals will develop Alzheimer’s Disease (AD) within the following five years; however several will remain stable and never develop AD (Gauthier et al., 2006; Petersen et al., 1999; Petersen et al., 2017). Although, there is no cure for AD, the early identification of individuals with MCI can enable treatments to delay the progression of the condition (Zucchella et al., 2018). Therefore, it is of paramount importance, to develop reliable objective diagnostic methods of cognitive impairment that can be conducted at primary care centers and memory clinics to determine whether an individual should seek further professional advice.

Methodology

90 individuals participated in the study. 23 SCI patients, 31 MCI patients and 36 healthy controls (HC) enrolled in the study. All participants were between 50 to 79 years old; had Swedish as their first and only language before 5 years old; had similar length of education; had no stroke or brain tumor; and had recent neuropsychological test results available for assessment. Connected speech data were elicited from cookie-theft picture description task (Goodglass & Kaplan, 1983), a standardized test employed in language therapy and evaluation sessions. Participants were recorded and the recordings were manually transcribed into text. The study refined the transcriptions of the recordings, defined several linguistic features, and employed two different annotation tools (Sparv and Parsey Universal) and two statistical measurements (Accuracy and Area under the Receiver Operating Characteristic (ROC AUC)) to select the superior feature set for the classification tasks. As a side product, an open source Swedish text annotation tool was deployed to benefit the linguistic research community. A novel feature engineering approach called SVC-Randomized Recursive Feature Elimination (SVC-RRFE) was introduced to select best features using Support Vector Machine, binary search and group *k*-fold cross validation. In the end, the 160, 150 and 98 selected features were applied and evaluated in feed-forward neural networks using group 10-fold cross-validation.

Results

Through group 10-fold cross validation neural networks (NN), we reached 76% mean accuracy, 73% mean ROC AUC, 0.47 mean Matthew’s correlation coefficient for MCI detection; 71% mean accuracy, 71% mean ROC AUC, 0.4 mean Matthew’s correlation coefficient for SCI detection and 75% mean accuracy, 71% mean ROC AUC, 0.39 mean Matthew’s correlation coefficient to differentiate MCI speakers and SCI speakers. The highest validation accuracy for the three models were 83%, 79% and 84%, respectively. The best features to classify MCI individuals and HC were mean length of word, words begin with [mɐ] and words with [ɪp] at the second and third position; the top 3 most important features to identify SCI individuals and HC were words with [ɑːg] at the second and third position, words begin with [mɑː] and words begin with [jøː]; and MLU, words begin with [dɛ], words with [ɑːd] at the second and third position were the most important features to differentiate MCI and SCI individuals.

Discussions

Phonology was impaired in patients with MCI and SCI subject. Specifically, Individuals with MCI showed more self-interruptions, produced more long vowels than the ones with SCI, more unrounded vowels than rounded ones and more stops follow by back vowels during the picture description task. Individuals with SCI tend to produce longer utterances than HC and MCI ones, and more nasal consonants follow by close front vowels. Sparv annotated data performed better during feature selection and the ones analyzed by Parsey Universal reached better results with neural networks. It proved that feed-forward neural networks can be used to build models to identify people with MCI and people with SCI. By employing phonological features this study provided improved classification of individuals with MCI, provided added objective markers than can be employed to identify these individuals for treatment.

Acknowledgments

I would like to thank my supervisor Haris Themistocleous for all the guidance, supervision and support throughout the project. He not only provided the idea of the project but also conducted the methodology of problem solving and paper writing.

I want to thank staff from Språkbanken, especially Dimitrios Kokkinakis, who provided me this opportunity and essential data to contribute to The Gothenburg MCI Study. Without his authorization of the data, this thesis would not be able to accomplish.

Furthermore, I would like to thank my parents who supported my studies in a spiritual and financial way. Thanks to my boyfriend for encouraging me to complete the project and take care of my life to assist the study.

Last but not least, I appreciate all reviewers for their suggestions and text improvements. Thanks for your effort!

Contents

Introduction

Mild Cognitive Impairment (MCI) is a neurological condition characterized by cognitive decline greater than expected for an individual's age and education level. Subjective Cognitive Impairment (SCI) is a self-reported decline in cognitive abilities but not clinically identified as MCI. Individuals with MCI remain functional in their daily activities (Petersen et al., 1999) and are characterized by different deterioration rates depending on the evaluation methods employed. More than 50% of these individuals will develop Alzheimer's Disease (AD) within the following five years; however, several will remain stable and never develop AD (Gauthier et al., 2006; Petersen et al., 1999; Petersen et al., 2017). Although there is no cure for AD, the early identification of individuals with MCI can enable treatments to delay the progression of the condition (Zucchella et al., 2018). Therefore, it is of utmost importance, to develop reliable objective diagnostic methods of cognitive impairment that can be conducted at primary care centers and memory clinics to determine whether an individual should seek further professional advice.

To identify indicators for diagnosis and prognosis of the condition, several studies applied natural language processing and signal processing on the language and speech of individuals with MCI, and linguistic features such as syntax, semantics, phonetics, phonology and discourse were analyzed (Beltrami et al., 2016; Fraser, Lundholm Fors, Eckerström, et al., 2019; Gosztolya et al., 2019, 2016; Roark et al., 2011; Themistocleous, Eckerström, et al., 2018). Combining acoustic information, individual meta-data and partial clinical test results, researchers managed to build language models on limited data in multiple languages (Beltrami et al., 2016; Gosztolya et al., 2019, 2016; Themistocleous, Eckerström, et al., 2018). However, many of them (Beltrami et al., 2016; Clark et al., 2016; Croot et al., 2000; Fraser et al., 2014; Fraser, Lundholm Fors, Eckerström, et al., 2019; Jarrold et al., 2014; Macoir et al., 2019; Pakhomov et al., 2010; Yu et al., 2015) relied on diagnostic batteries and gathered the features by manual counting, which were time consuming since the participants need to take multiple testing in clinic, and the process of visiting clinical facilities and preparing test environment are complicated. Further, the collection and organization of the result of batteries need to follow different criteria which made the whole activity hard to quantify. While most mentioned related findings focus on people with MCI or AD, very few studies provide evidence from SCI (Linz et al., 2019; Lundholm Fors et al., 2018; Macoir et al., 2019; Rabin et al., 2017) . Although language models built with known predictors show high accuracy, phonology production of MCI and SCI individuals had not been studied extensively. Looking up the models with top accuracy and ROC-AUC in the recent studies to detect individuals with MCI from healthy control, Fraser et al. (2019) gained 83% accuracy, 80% sensitivity, 85% specificity and 88% ROC-AUC with language, speech, eye movement and comprehension features; and Gosztolya et al. (2016) reached 88.1% accuracy, 88.5% unweight average recall and 89.1% F-score with acoustic features and metadata of the participants (gender, age, education, etc.). Although the work by Gosztolya et al. (2016) investigated phonology of MCI patient, the markers were extracted on acoustic level rather than text level, and the phonological impairment of MCI subjects were not discussed. Therefore, this thesis work aimed at identifying the phonological phenomenon of MCI and SCI subjects, and the following questions were asked:

1. Do individuals with MCI/SCI have significant differences in phonology compared to normal ageing in Swedish speech?
2. Do individuals with MCI have significant differences in phonology compared to SCI in Swedish speech?

Specifically, the goal of the work is to determine linguistic boundaries that help with MCI and SCI detection and to determine novel features that classify MCI vs SCI, MCI vs HC and SCI vs HC through natural language processing and machine learning techniques. We analyzed speech transcriptions of Swedish individuals without additional information from neurological tests and applied different annotation tools and statistical measurements to determine the best language models. During the study, an open-source tool was deployed for Swedish text annotation based on Google's Syntaxnet (Alberti et al., 2017) language models. The tool is portable and could be configured for other language exploration tasks. To extract phonological information from the given data, studies of its context and distribution were conducted. The pattern of phonemes was observed based on refined transcriptions of speech. A few linguistic predictors from previous studies were selected and marked by annotation tools. Two annotation tools were employed, Sparv and Parsey Universal. The former was developed in academic context and only for Swedish text, the latter came from the IT industry and was developed for all languages. Since both tools support Swedish text annotation, we employed them respectively to pursue the best performance of the final classification. Comparing the validation result of the models trained on the features extracted by the annotation tools, we determine which annotation tool is more suitable for the present study. Another highlight of the study is the feature elimination strategy. To reduce the dimensionality of the data, we conducted dimensionality reduction after evaluating dimensionality reduction methods. It provides an idea to improve existing feature selection algorithms and could be applied on other dimensionality reduction scenarios. A novel feature selection method called Randomized Recursive Feature Elimination (RRFE) was proposed and Support Vector Classifier (SVC) was used to evaluate the weight of eliminated predictors. Compared to common Support Vector Classifier-Recursive Feature Elimination (SVC-RFE) strategy, it uses binary search to reduce redundant predictors and turn out faster in finding the best feature set under limited experiment setup (i5-6200U CPU, 8GB RAM, 2G GPU NVIDIA GeForce 920M). Accuracy and Area under the operating characteristic Curve (ROC-AUC) were employed and compared as benchmark during feature selection, combined with Matthews correlation coefficient. Twelve binary classification experiments were conducted to determine the best predictors for the classification tasks. Feed forward neural networks were employed for the classification tasks. Also, we evaluated different types of classification models based on their performance to classify patients with MCI from healthy controls, MCI patients from SCI patients, and SCI patients from healthy controls. Our findings directly contribute to diagnosis of MCI and SCI from a linguistic perspective in Swedish individuals. To the best of my knowledge, this is the first work that investigated the phonemic production of SCI patients and applied it to distinguish them from normal ageing. It is as well the first one that examined the phonemic pattern of Swedish individuals with MCI and SCI. It also proved that detection of neurological conditions could be much cheaper and more efficient through machine learning technologies.

Background

MCI, SCI and present diagnosis methods

The severity of MCI has been associated with its conversion to dementia and other cognitive conditions. Although individuals with MCI were not observed with massive difficulty in daily activities(Petersen et al., 1999), Budson & Solomon (2011) found that 70% of diagnosed MCI subjects progressed to dementia and Farias et al (2009) claimed that the annual conversion of MCI to dementia in clinic samples often range from 10% to 15%. According to the World Health Organization Dementia Fact Sheet, Alzheimer’s Disease (AD) is the most common type of dementia which covers 60%-70% of the cases. It not only causes disability and dependency among elderly people, but brings high medical, social care and informal care costs to society (Brodaty & Donkin, 2009). Although AD is currently incurable, diagnosing at its earlier stage (like MCI) could provide patients and caregivers longer time to adjust life plan, optimize support, detect and treat challenging symptoms and complications, and delay potential loss.

The relation between SCI and MCI was discussed in several studies. Jessen et al. (2014) found that people with SCI could progress to MCI after six years, and those who confessed having memory decline showed impairment in remembering specific events on a follow up test after eight years observed by Koppara et al. (2015). On the contrary, a few findings implied that SCI might not lead to MCI or AD (Garcia-Ptacek et al., 2014) and could be related to or induced by emotions like depression and anxiety (Balash et al., 2013; McDougall et al., 2006; Yates et al., 2017). However, depression and anxiety could also be found on MCI subjects (Barnes et al., 2006; Bhalla et al., 2009), which indicated that SCI might be the earliest stage of AD progression before MCI and Reisberg & Gauthier (2008) suggested that identification of SCI could potentially reduce the risk to develop to AD.

MCI in clinical assessments is diagnosed through neuropsychological testing, neuroimaging (Positron Emission Tomography (PET), structural magnetic resonance imaging (MRI) (Studart & Nitrini, 2016), diffusion tensor imaging (DTI)) and electroencephalogram (EEG) (Alberdi et al., 2016). Neuroimaging was also applied on SCI studies (Kryscio et al., 2014; Yasuno et al., 2015) as well as questionnaires and scales. However, there is a lack of standard with balanced categorization, proper time period, perception of subjects and questionnaire design. Memory complaints, cognitive function declines, generic issues and specific daily events were questioned in the surveys and a few episodic memory tests were created to detect the correlation between pre-clinical stage of AD and SCI (Studart & Nitrini, 2016). Besides questionnaire and scales, Mini-mental State Examination (MMSE), delayed recall test and Dementia Rating Scale-Mattis (DRS-Mattis) were applied for SCI detection from individuals (Aguiar et al., 2010; Brucki & Nitrini, 2009).

Language study on MCI and SCI

Language is severed in AD and its early stages (Szatloczki et al., 2015). While traditional clinical methods to diagnose AD and its early stages are straight-forward, the complexity, risk factor and expense of them are not ideal to the majority (Clark et al., 2016). Language assessments provide a simpler, economic and reliable approach to physiological methods. Since individuals with MCI have memory problems and language skills require memory functions, there could be impacts to their language ability and the observed pattern could be employed as predictors. Clark et al. (2016) found that patients with MCI are impaired in word similarity and verbal fluency tasks. Also, analysis of eye-movement during text reading showed that individuals with MCI were characterized by disorganized and non-linear reading path (Fraser et al., 2017). Hernández-Domínguez et al. (2018) found smaller vocabulary and less complexity of syntax in cognitive impairment. While linguistic functions were widely investigated on MCI subjects, study on distinguishing SCI is very rare at only linguistic aspects (Rabin, Smart, & Amariglio, 2017; Linz et al., 2019; Macoir, Lafay, & Hudon, 2019). Lundholm Fors et al. noticed that SCI subjects tend to use longer and more complex expressions in speech compared to healthy control and MCI subjects. Macoir, Lafay, & Hudon observed that SCI subjects are impaired in verbal fluency at midpoint between HC and MCI. Overall, all findings show that people with SCI behave very similar to normal ones in speech and no significant linguistic boundary to differ them.

The traditional ways to identify language impairment in people with neurological conditions are diagnostic batteries, such as the Boston Naming Test (BNT) and Western Aphasia Battery (WAB). Both evaluation batteries test linguistic and non-linguistic skills where the latter tests reading, writing, drawing and calculation skills as well. Diagnostic batteries were proven to be reliable and have been employed in related studies broadly. However, they require a lot of manual labor, to collect and score the outcomes. Since speaking is a daily cognitive activity in most cases, it takes much less effort to collect information from subjects. Automatic language analysis, using state-of-the-art speech recognition technologies (Konig et al., 2018) and natural language processing can provide objective diagnostic markers that are easier and faster to elicit.

Language of patients with MCI using natural language processing

Morphological markers

Several studies employed morphological features. For instance, Vincze et al. (2016) showed that function words, the ratio of adjectives, ratio of pronouns, ratio and count of conjunctions were the critical morphological markers of patients with MCI. Another study found that the frequency of verbs is a statistically significant feature that distinguishes patients with MCI from HC in classification tasks (Beltrami et al., 2016). In a recent classification study of patients with MCI and HC, Kathleen C. Fraser et al. (2018) found that the proportion of main clauses with nonfinite or finite verbs, counts of nouns, verbs, noun-verb ratio and verb frequency were one of the highly ranked predictors; in a small size of corpus, MCI participants showed reduction of nouns, slight increase of verbs, lower noun-verb ratio and higher verb frequency. Based on the literatures, we suggest the first study hypothesis that POS morphological features distinguish individuals with MCI from SCI, and HC.

Study Hypothesis 1: POS morphological features distinguish individuals with MCI from SCI, and HC.

Syntactic markers

Syntactic complexity has been investigated on different neurological conditions in previous studies but giving conflicting findings. A recent assessment (CAN & Gülmira, 2018) showed that early-onset AD patients tend to produce shorter, less verbal and less complex sentences in speech in terms of coordinated or compound sentences. On the other hand, a few studies concluded no significant syntactic difference between AD and HC (Kave & Levy, 2003; Wei et al., 2018) A study showed *that syntactic factors varied among different patients with dementia* (Wei et al., 2018), and similar outcome from Kave & Levy proved that AD patients express less information with more semantic mistake but same syntactic structures and morphological forms in speech. When identifying MCI subjects, Lundholm Fors et al. (2018) found no significant syntactic features that could differentiate MCI and HC group, and MCI and SCI group. Under this context, we propose the second study hypothesis that syntactic complexity measures distinguish individuals with MCI from SCI, and HC.

Study Hypothesis 2: Syntactic complexity measures distinguish individuals with MCI from SCI, and HC.

Phonological markers

While morphology and syntax were very much investigated in MCI and AD, phonological features were comparably less explored (Croot et al., 2000; Meilán et al., 2012; Themistocleous, Eckerström, et al., 2018; Yu et al., 2015). Many related studies on cognitive impairment focused on acoustic and speech predictors such as articulation rate, speech tempo, hesitation ration, silent pause etc. (Roark et al., 2011; Satt et al., 2013; Tóth et al., 2018, 2015). Croot et al., (2000) noted that phonological impairment may occur in the early stage of AD where patients showed difficulty to access phonological form when naming pictures and tend to make errors when initializing sounds in words. Meilán et al. (2012) found that voiceless segments produced by AD patients were highly correlated with their phonological fluency. Computational methods were applied in recent studies to investigate language impairment of MCI individual at phoneme level. Themistocleous, Kokkinakis, et al. (2018) observed that Swedish MCI individuals produce longer vowels than healthy controls during text reading tasks. Yu et al. (2015) built an SVM classifier with total duration of '*s*' phoneme, pseudo-syllable rate, average pause duration, total count of '*m*' phoneme for automated cognitive impairment diagnosis. A phonetic feature selection study for MCI subject detection by Gosztolya et al. (2016) combined the number of occurrences, mean length and standard deviation of the occurrences of given phonemes. Their findings showed that phoneme-related predictors contributed 40% of the best feature set selected from 235 candidates. Based on the related studies, we propose that phonological features distinguish individuals with MCI from SCI, and HC.

Study Hypothesis 3: Phonological features distinguish individuals with MCI from SCI, and HC.

Machine learning on MCI and SCI

Earlier studies employed computational linguistic technologies to diagnose patients with MCI and AD in different languages. For example, Asgari et al. (2017), Fraser et al. (2013, 2016), Orimaye et al. (2017), Orimaye et al. (2014) and Roark et al. (2011) explored linguistic changes on different stages of cognitive impairment and dementia in English speakers. Gosztolya et al. (2019, 2016) and Vincze et al. (2016) investigated multiple linguistic markers to identify individuals with MCI, mild AD from normal ageing in Hungarian. Beltrami et al. (2016) managed to collect transcriptions of Italian MCI individuals with picture and topic description tasks. Santos et al. (2017) tried to find a general method to distinguish MCI individuals and healthy control in English and Portuguese corpus. In Swedish, Themistocleous et al. (2018), Lundholm Fors et al. (2018, 2017) examined acoustic and linguistic features based on data collection of Kokkinakis et al. (2017). Fraser et al. (2019) also analyzed the same corpus of Swedish MCI participants using multiple language skill evaluation tasks including material of reading silently, reading aloud, and picture description tasks. Fraser, Lundholm Fors, & Kokkinakis (2019) combined multiple data collections from English and Swedish speakers, extracted word embeddings of nouns and verbs and distinguished the ones with MCI from healthy control through unsupervised topic modelling clusters.

Machine learning models can facilitate automated diagnosis and prognosis of dementia in AD and MCI. A commonly employed learning algorithm to detect individuals with MCI is support vector machine (SVM), which is a supervised learning model that finds a hyperlane to separate vectorized samples into different groups. Among 7 out of 12 recent related works that employed machine learning, Fraser et al. (2019) got its best model with 83% accuracy, 80% sensitivity, 85% specificity and 88% ROC-AUC; In another multi-language study, Fraser, Lundholm Fors, & Kokkinakis (2019) achieved 72% accuracy, 77% sensitivity, 67% specificity on Swedish speakers, and 63% accuracy, 53% sensitivity, 74% specificity on identifying English speakers with MCI. Gosztolya et al. (2019, 2016) reached 86% and 88% accuracy with 5-fold and 10-fold cross validation SVM trained classifiers on MCI vs HC detection of the participants, and 80% accuracy on distinguishing MCI and mild AD ones. Asgari et al. (2017) achieved 83% accuracy, 81% sensitivity, 76% specificity and 80% ROC-AUC with 5-fold cross validation SVM, Santos et al. (2017) got 75% accuracy, Vincze et al. (2016) reached 75% accuracy and Roark et al. (2011) reached 73.2% ROC-AUC on MCI detection of the individuals.

While most promising models were built by SVM, neural networks, random forest and Naive Bayes were also explored to detect individuals with MCI. Lundholm Fors et al. (2018) applied leave-one-out cross validation using random forests with only syntactic features and reached a 68% F-score on MCI individuals vs HC, 66% F-score on MCI vs SCI and 54% F-score on SCI individuals vs HC. Fraser et al. (2017) reached 84% accuracy on participants with MCI vs HC detection based on Naive Bayes. Themistocleous et al. (2018) achieved 83% accuracy by only acoustic features and individual meta-data using 10-hidden-layer feed-forward neural networks.

Table 1 contains an overview of related studies that used machine learning techniques and natural language processing on cognitive impairment characterization. As the table shows, many factors affect the performance of the language models, such as data collections (quantity and quality of the data), feature selections (linguistic and non-linguistic information), experiment design, classifier selections, etc. Evaluations of model performance differ from various studies. Although the collection is not exhaustive, it presents a variety of methods and the significance of investigating linguistic functions on MCI and SCI subjects. It also reveals the linguistic markers that are influenced during pathology of the conditions.

Table 1: Studies on cognitive impairment characterization with linguistic features and machine learning technologies for the classification of patients with Mild Cognitive Impairment (MCI), Mild Alzheimer's Disease (Mild AD), Subjective Cognitive Impairment (SCI) and healthy controls (HC)

Problem	Classifier	Feature	Data collection	Best Performance	Research
MCI vs HC	Leave-pair-out cross validation Logistic regression and Support Vector Machine (SVM)	Speech features, language features, eye features, comprehension features	The Swedish Cookie-Theft Corpus	83% accuracy, 80% sensitivity, 85% specificity, 88% ROC-AUC with all features from all tasks by SVM	Fraser et al. (2019)
MCI vs HC	Leave-one-out cross validation SVM	Features of unsupervised topic modelling cluster generated by word embeddings of nouns and verbs.	The Swedish Cookie-Theft Corpus (Swedish), Karolinska Corpus (Swedish), Dementia Bank Corpus (English)	Swedish: 72% accuracy, 77% sensitivity, 67% specificity English: 63% accuracy, 53% sensitivity, 74% specificity	Fraser, Lundholm Fors, & Kokkinakis (2019)
MCI vs HC MCI vs Mild AD (mAD)	5-fold cross validation SVM	Acoustic features, morphological features, dialog events, semantic features	Hungarian MCI-mAD database	MCI vs HC: 86% accuracy, 86% F-score with acoustic features and semantic features MCI vs mAD: 80% accuracy, 81.5% F-score with acoustic features and all linguistic features	Gosztolya et al. (2019)
MCI vs HC	Neural networks	Acoustic features and individual metadata	The Swedish Cookie-Theft Corpus	83% accuracy with all features	Themistocleous, Eckerström, et al. (2018)

MCI vs HC; MCI vs SCI; SCI vs HC	Leave-one-out cross validation random forest	Automated syntactic features (dependency distance, phrase type proportion, phrase type length)	The Swedish Cookie-Theft Corpus	MCI vs HC: 68% F-score MCI vs SCI: 66% F-score SCI vs HC: 54% F-score	Lundholm Fors et al. (2018)
MCI vs HC	Naive Bayes	Eye-movement features, word frequency, part of speech tags from text reading task	The Swedish Cookie-Theft Corpus	84% accuracy with all features	Fraser et al., (2017)
MCI vs HC	5-fold cross validation SVM	Morphological features, lexical features grouped by word sense, and dialogue events.	English unstructured conversation from interview with preselected topic across subjects	83% accuracy, 81% sensitivity, 76% specificity, 80% ROC-AUC with 10 selected features (verb tense and verbs about motion)	Asgari et al. (2017)
MCI vs HC	SVM	Statistics of word adjacency model enriched by word embeddings, linguistic features and bag of words.	The Cookie-theft picture description dataset (English), The Cinderella narrative dataset (Portuguese), The ABCD Dataset (Portuguese)	Cookie theft: 65% accuracy with all features extracted from raw transcriptions. Cinderella: 65% accuracy with statistics of word embedding enriched word adjacency model extracted from raw transcriptions. ABCD: 75% accuracy with bag of word extracted from raw transcriptions. Cinderella: 72% accuracy with linguistic features extracted from manual processed data.	Santos et al. (2017)

MCI vs HC	SVM	Morphological features, semantic features, dialog events, patient metadata.	Hungarian speech transcription on two short animated films	75% accuracy with 35 selected morphological, semantic and dialogue features	Vincze et al. (2016)
MCI vs HC	10-fold cross validation SVM	Phonological features, acoustic features, patient metadata	Hungarian MCI-mAD database	88.1% accuracy, 88.5% Unweighted average recall, 89.1% F-score	Gosztolya et al. (2016)
MCI vs HC	Neural networks	Speech features, lexical features, syntactic features	Italian speech transcription on three description tasks (picture, working day, dream)	76% accuracy, 76.7% precision, 75.4% recall, 75.9% F-score with 17 selected features	Beltrami et al. (2016)
MCI vs HC	SVM	Linguistic feature (words per clause, syntactic feature, POS tag cross entropy, content density), speech feature (Pause rate, phonation rate etc.)	English audio and speech transcription on three-sentence story recall task	70.3% ROC-AUC with 5 selected linguistic features and 4 selected speech features; 73.2% ROC-AUC with 5 selected linguistic features	Roark et al. (2011)

Material and Methodology

To distinguish individuals with MCI, SCI and healthy controls (MCI vs SCI, MCI vs HC, HC vs SCI), feature selection and model comparison with machine learning were performed on the data. Feature selection was conducted in two stages, which includes feature collection and feature elimination. The candidate predictors were collected based on the findings of related studies. A pre-study was performed to determine the most suitable algorithm for feature selection. Compared with the accuracy of using Naïve Bayes and Decision Tree for the classification tasks, Support Vector Machine was chosen to select the best feature composition. Feed-forward neural networks were adopted to verify the performance of selected features, and prediction accuracy, Matthews correlation coefficient (MCC), area under the receiver operating characteristic curve (ROC AUC) value were applied as benchmarks to evaluate the models. Group k-fold cross validation was deployed in both feature selection and neural networks to avoid training and validating on the same individuals. The following chapter describes the procedure of the present study in sequence, which includes study design, feature selection, model building and training. Data acquisition, annotation tools, and predictor collection were explained in study design. The prediction methods and feature elimination are discussed in feature selection. The method of model building and training was described in the last section with the explanation of evaluation methods.

Study design

Data source

The data comes from the Gothenburg MCI Study (Wallin et al., 2016), which is a large scale longitudinal in-depth phenotyping study of patients with different forms and degrees of cognitive impairment using neuropsychological, neuroimaging, and neurochemical tools. It is based on clinical tests and aims at improving the detection of early and manifest stages of dementia and its complications. The Gothenburg MCI Study is approved by the local ethical committee review board (reference number: L091-99, 1999; T479-11, 2011). The currently described study is approved by the local ethical committee (decision 206-16, 2016).

In this work, we investigated speech transcriptions from 90 individuals: 23 SCI patients, 31 MCI patients, and 36 healthy controls. Participants were selected following the inclusion and exclusion criteria (Kokkinakis et al., 2017), naming that the participants should be:

- between 50 to 79 years old,
- have Swedish as their first and only language before 5 years old,
- have similar length of education,
- no stroke or brain tumor, and
- have recent neuropsychological tests result available for assessment.

Moreover, they should not have any underlying mental condition, poor vision, or hard understanding of the context of data collection.

Speech recordings were elicited from the Cookie-theft picture task (see Figure 1) from the Boston Diagnostic Aphasia Examination (Goodglass & Kaplan, 1983). This is a picture showing two children trying to steal cookies from a cookie jar “KAKBURK”. The Cookie-theft picture is commonly employed for eliciting speech for cognitive impairment analysis. It is especially beneficial for research result comparison on cognitive impairment since many other data collections were acquired with it (Kokkinakis et al., 2017).

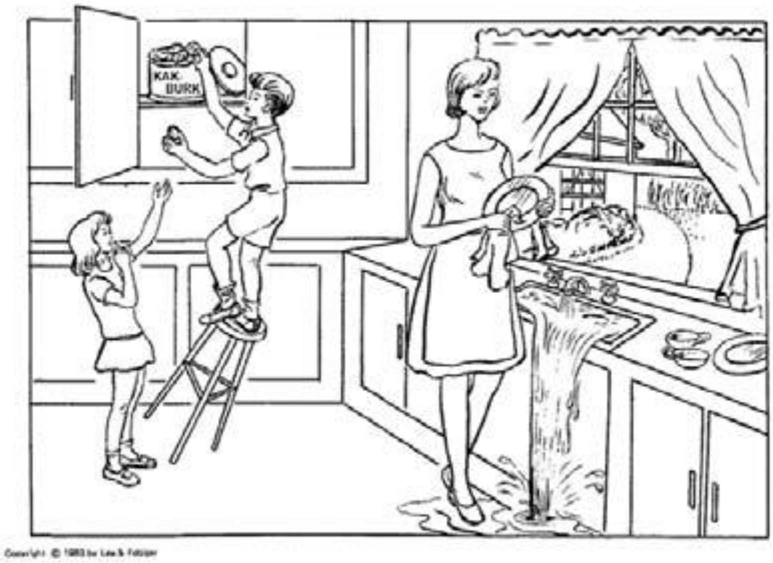

Figure 1: The Cookie-theft picture

During recording, participants were asked to describe observed objects and events from the picture as much as possible without disturbance. The audios were captured using an H2n Zoom recorder (wav files, 44.1 kHz with 16-bit resolution without compression). The audio recordings were transcribed with PRAAT application (Boersma & Weenink, 2020) using 2-tier text grid configuration which contained both standardized spelling and maintained spontaneous speech form. The later form included partial words (self-interruption), filled pauses (hesitation, repetition, etc.) and non-verbal sound (laughter, coughing, etc.).

Annotation tools

To answer which annotation tool is better for the present classification tasks, morphological and syntactic features were extracted by Parsey Universal (Alberti et al., 2017) and Sparv (Borin et al., 2016), respectively. Parsey Universal is a component of Google's Syntaxnet and Sparv is a Swedish annotation tool developed by Språkbanken of University of Gothenburg (Borin et al., 2016). Both annotation tools analyze transcribed text not acoustic data.

Data processing through Parsey Universal was deployed by a local flask-built API service which returns POS tags of words of the given sentence upon request. Inspired by andersrye/parsey-universal-server (Rye, 2016/2019), the service was set up in Python using Flask (Grinberg, 2014) and packed into a docker container so that it could be installed and accessed by anyone anywhere. It loads Google's Parsey Universal pre-trained syntactic model for any language that is available, tokenize and parse the sentence, with return of syntactic relation between words and their POS tags stored in json format. It has a simple graphic interface developed in html, where users can access and try parsing any Swedish sentences. As a sub product of this thesis work, the source code of the Parsey Universal server is now open to the public. Detailed instructions on how to deploy can be found through jesdoit/parsey-universal-server (*Jesdoit/Parsey-Universal-Server*, 2018/2018) on GitHub.

Sparv is a powerful Swedish corpus annotation tool which can analyze text on both word/token level and document level (Borin et al., 2016). It could manage lexical analysis, compound analysis, part-of-speech tagging, syntactic analysis and named entity recognition. Annotation by Sparv using its web interface. Since it has a user interface where people can upload full text for annotation directly, it is much faster and very easy to use for a big batch of data compared to Parsey Universal. By tailoring segmentation configuration with uploaded corpus, it returns annotation result in XML format which is another human and machine friendly data structure. Moreover, it is very convenient to use since the service and project data are provided from the same provider.

Data refinement, dialog events and POS features

The maintained spontaneous speech forms (see Table 2) were intentionally marked to retain the individual information from the speech, but on the other hand, could degrade the performance of text processing of annotation tools that are built for orthographic natural language processing. Therefore, data refining took place to avoid faulty segmentation, tokenization, and Part of Speech (POS) tagging.

Table 2 demonstrates the special symbols marked in the Cookie Theft transcripts. The original form was written in Swedish (the text was translated by the author to English).

Symbol	Example	Meaning
-	the-	Mark as started word that got interrupted
.	There is a vase on the table.	Mark as sentence boundary
--	He is standing close to -- close to the table	Mark as interrupted sentence opposite to written language
<PAUS></PAUS>	<PAUS>silence</PAUS>	Mark as break
<SKRATT></SKRATT>	<SKRATT> </SKRATT>	Mark as laughter
<ANNAT></ANNAT>	<ANNAT>coughing</ANNAT>	Mark as other sound
<OHÖRBART></OHÖRBART>	<OHÖRBART>x</OHÖRBART>	Mark as unheard syllable

The data was originally categorized into three folders for HC, MCI and SCI. Each folder contains a number of transcriptions in txt format named by the anonymous codes of individuals. The individual texts consist of the name of the audio file and transcribed sentences segmented by newlines. We extracted the label and the individual for each sentence, shuffled them, and stored them as Pandas DataFrame (McKinney, 2010), which is a two-dimensional data structure that stores data in rows and columns in Python. Special tags like "<ANNAT>", "<OHÖRBART>", "<SKRATT>", "<PAUS>" represent "other", "unclear", "laugh" and "pause", were removed to link up sentences. Events like repetition, self-correction, interruption and unclearly pronounced terms that happen very often in natural discourse were also considered. Several rules were defined to capture the frequency of the events with regular expression. Duplicated terms that could mess up with POS tagging were removed as well. The frequencies of different events were saved in a csv file for further process, along with refined transcription.

The following rules applied for the present transcription:

- Replace all "<OHÖRBART>" to "NÅT" (that means "something" in English), mark **unclear+1**
- Remove all "<ANNAT>"

- Remove all “<SKRATT>” since this tag only appears twice and on the same individual.
- Replace “-” in common words. For example, “t-shirt” to “Tshirt”.

For each sentence with “-”:

- if “-” appears in the end of word:
    - if the word before “-” equals to the one right after, or, two words that before and after it are the same, remove “-” word and the one before it from the sentence, mark **self-correction+1**, **repeat+1**
    - if the word with “eller” (means “or” in English) right after it, remove only “-”, mark **self-correction+1**
    - if both words before and after it are any of “i” (“in” in English) or “på” (“at” in English) or “som” (“as” in English), remove the word and the one before it, mark **self-correction+1**, **Preposition-error+1**
    - if the letters before “-” of the word are part of the word after it, remove “-” word, mark **self-correction+1**, **repeat+1**
    - if no rules above fulfils, remove “-” word, mark **interruption+1**, **self-correction+1**
- if “-” appears in the middle of a word:
    - if the letters before “-” equals to the ones after, remove the letters before “-”, mark **self-correction+1**, **repeat+1**
- If “-” appears at the end of the sentence:
    - remove it and mark **self-correction+1**, **interruption+1**

The original size of refined data is 1316 which contains 505 healthy control (HC), 442 mild cognitive impairment (MCI) and 369 subjective cognitive impairment (SCI) samples. The amount of HC data is about 27% more than SCI data which may bring bias during HC vs SCI classification. Therefore, we group the samples into three by downgrading healthy controls to the size of MCI ones. The HC samples were selected by a fixed random state value (42) to get the experiment results reproducible. The resampled data for three classification tasks is 442 HC vs 442 MCI, 442 HC vs 369 SCI, and 442 MCI vs 369 SCI, in respect to Figure 2.

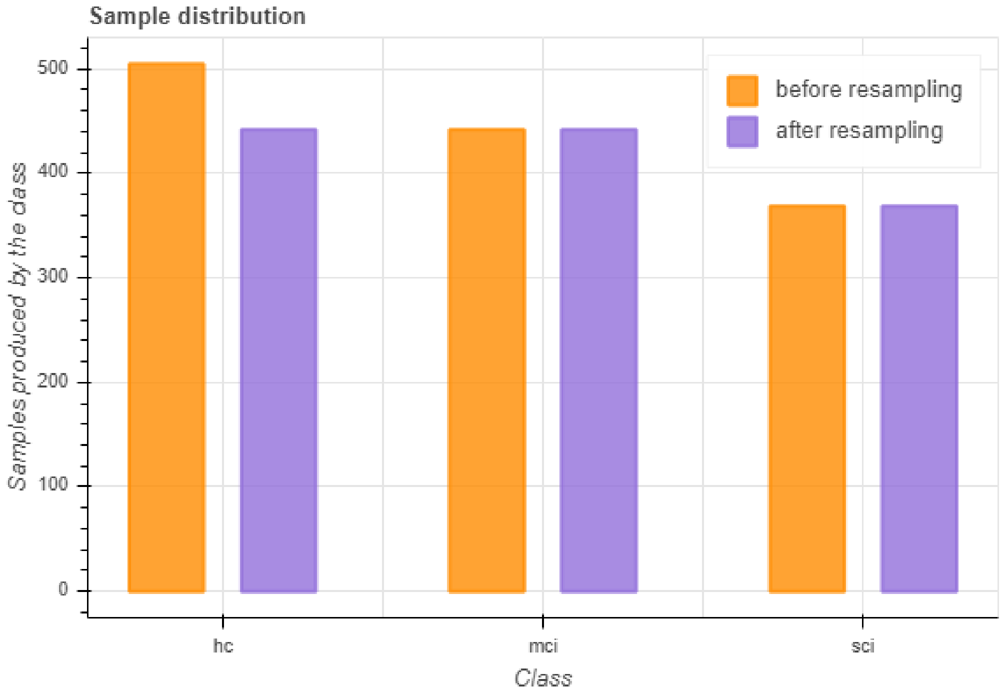

Figure 2: The data distribution before (in orange) and after (in purple) resampling. The x-axis refers to the class of the data (hc=healthy control, mci=individual with mild cognitive impairment, sci= individual with subjective cognitive impairment) and the y-axis shows the number of samples produced by the class.

All POS tags of words were stored into lists to generate the log frequency of nouns, verbs, light verbs, determiners, prepositions, adjectives, adverbs, pronouns, conjunctions, particles, participles and the infinitive words of each sentence. In addition to POS counts, log noun-verb ratio was computed by dividing count of nouns with count of verbs. All values were rounded to two decimals.

Syntactic Features

To verify whether syntactic complexity is helpful for our classification tasks, we made use of the dependency tree from the annotation result which contains syntactic information. The selection of syntactic complexity markers was motivated by (Kokkinakis et al., 2017), which suggested investigating dependency distance on the same data source. Therefore, the average dependency distance (ADD) shown in (1) and the average verb distance (AVD) (2) were measured. Dependency distance is measured as the gap between a specific word and its dependency head. For each sentence, we calculated dependency distance on all words and divided them by the number of words to retrieve the average value. The concept of verb distance is to observe the gap between all head nodes (verbs) from the dependency tree. Average verb distance is then computed by dividing total verb distance of the sentence with count of verbs. All values were rounded to two decimals.

$$ADD = \Sigma \frac{|w_i - w_{parent\ of\ i}|}{count\ of\ w} \qquad (1)$$

$$\text{AVD} = \frac{\Sigma|verb_i - verb_j|}{count\ of\ verb} \qquad (2)$$

The other selected syntactic indicators are mean length of utterance (MLU) and mean length of word (MLW) which were investigated in Fraser et al. (2014). They were proved to be key features to distinguish semantic dementia subjects from healthy controls. The MLU shown in (3) of each individual was calculated by summing up the length of each refined sentence from the same individual divided by the count of his/her total speech. MLW show in (4) was measured with the average length of words counted by letters for each individual. The values were rounded to two decimals and distributed back to each sentence in respect to the individual.

$$MLU = \frac{\Sigma(count\ of\ word\ in\ sentence)}{count\ of\ sentences\ by\ speaker\ X} \qquad (3)$$

$$MLW = \frac{\Sigma(length\ of\ word)}{count\ of\ words\ by\ speaker\ X} \qquad (4)$$

Phonological Features

To generate phonemic features from refined samples, python module Epitran v0.46 (Mortensen et al., 2018) was introduced to transcribe orthographic Swedish text into International Phonetic Alphabet (IPA) format.

The set of phonemes and vowels were extracted from transcribed IPA text. Positional segment frequency, biphone frequency, average length of syllable and average syllable complexity were selected as phonological candidate predictors for feature selection.
The approach of positional segment frequency (PSF) (5) and biphone frequency were introduced in Storkel's work (Storkel, 2004b). For every phoneme at a specified position in the dataset, the PSF was calculated by log frequency of it in each word of the sentence, divided by all phonemes' log frequency in each word of the sentence. To avoid division by zero, one (1) is added after log of counts. For instance, to calculate PSF of phoneme "*ɛ*" at second index of word in sentence "Hello World" [h.ɛ.l.əʊ w.ɜː.l.d], the equation is:

$$PSF = \frac{\log(count\ of\ \text{ɛ}\ at\ second\ position) + 1}{\Sigma(\log(count\ of\ [h, \text{ɛ}, l, \text{əʊ}, w, \text{ɜː}, d] at\ second\ position) + 1)} \tag{5}$$

Since [*h*], [*l*], [*əʊ*], [*w*], [*d*] appear 0 time at the second index of any word, their log frequencies in this case are all 0. The result of PSF calculation should be 0.5, by

$$\frac{log1 + 1}{0 + (log1 + 1) + 0 + 0 + (log1 + 1) + 0} = 0.5 \tag{6}$$

The difference between PSF and biphone frequency is that the latter takes bigram phoneme into calculation instead of only one phoneme. That is, to know the biphone frequency of [*hɛ*] at first index of the word in sentence "Hello World", you have:

$$\frac{log1 + 1}{(log1 + 1) + (log1 + 1)} = 0.5 \tag{7}$$

Where among all different bigram combinations of phonemes in the example, only [*hɛ*] and [*wɜː*] are at first index of word in the sentence. The rest have no log frequency.

Following the concept, the first and second index of PSF and biphone were considered in this thesis work. The result of calculations was padded into the same dimension with 0 for ease of further process. The total number of valid PSF and biphone frequency predictors was 619, based on 35 phonemes and 1225 pairs of biphone.

In addition to positional segment frequency and biphone frequency, average syllable length and average syllable complexity were also included. Syllable complexity was correlated with two-word complexity measurement back in 1954, where Menzerath claimed that the increase of syllables per word leads to the decrease of phonemes per syllable. Measure of numbers of syllables in speech content has been proved as an effective indicator to distinguish aphasia from HC with an updated version of the "Cookie Theft" picture (Shauna et al., 2019).

Average syllable length (ASL) (see 8) equals the sum of vowel frequency in each transcribed word of the given sentence divided by the length of sentence. Average syllable complexity (ASC) (see 9) is measured by dividing the sum of syllable complexity of each word with length of sentence, where syllable complexity of word is obtained by calculating the length of word with at least one vowel divided by its vowel frequency.

$$ASL = \frac{\Sigma(count\ of\ vowel)}{count\ of\ words\ in\ sentence} \tag{8}$$

$$ASC = \frac{\Sigma\left(\frac{count\ of\ phoneme\ in\ word}{count\ of\ vowel}\right)}{count\ of\ words\ in\ sentence} \tag{9}$$

Feature selection

The sum of candidate predictors was 689 which includes all ones described in the preceding section. The high quantity of features showed poor performance with alternative machine learning algorithms in this study. Algorithms like Naive Bayes, Decision Tree, Support Vector Machine from scikit-learn 0.19.0 (Pedregosa et al., 2012), and Feed Forward Neural Networks were estimated during classification method pre-study. None of them reached higher than 45% accuracy on validation data with 689 features and it was expensive to train a model with so many features thinking of time consumption. Thus, feature selection technology was employed to improve the performance of classifiers.

Prediction methods

Since the size of the resampled data is small for machine learning, it is beneficial to get use of all available samples in model training. Therefore, the k-fold cross validation method was adopted during prediction. It splits the dataset into k smaller ones, recursively trains the model on k-1 sets and validates it with the remaining one. The performance of cross validation is measured by averaging the results of each rotation. A key for k-fold cross validation in this project is splitting data by group information based on the owner of the speech samples. Group k-fold cross validation is recommended for domain specific data, such as multiple medical samples from multiple patients. Since our data contains highly individual information, group k-fold is the best practice to prevent models from training and validating on the same individual. The GroupKFold class from scikit-learn does the desired technique. It creates a specified number of train test sets and avoids the same individual’s transcription appearing in both ones. The value of k was selected based on the number of data points. For the HC vs MCI group, 4, 6, 8 and 10-fold were tested to find features that could provide the highest accuracy or ROC AUC value. For the other two groups, 5, 6, 8, 10, 12-fold were tested to select the features with the highest accuracy or ROC AUC.

All data were standardized. Forming feature value in a certain range is important when it contains varied information calculated in different ways. MinMaxScaler was selected among all other scaling methods in scikit-learn such as StandardScaler, MaxAbsScaler and Normalizer. By scaling feature values to [0, 1] range, it gives comparably better performance combined with support vector machine approach. Only the training set was fit by MinMaxScaler before transforming so that testing set could apply the same shifting operation without revealing the information from training data.

Feature elimination

A paper titled *Irrelevant features and the subset selection problem* authored by John et al. (1994) inducted feature selection methods into two main categories: wrapper and filter methods. Wrapper methods are search algorithms that conclude features for optimizing model performance by adding or removing predictors with classification algorithms. Filter methods only keep the features that are highly relevant with the target problems. Since it was unknown which features are highly correlated with MCI or SCI individuals, we could not go for the filter method with limited domain knowledge.

A wrapper method called Randomized Recursive Feature Elimination (RRFE) was instead employed to extract the most optimal features with binary search and SVM classifier. The idea of this approach is inspired by Guyon et al. (2002) where they introduced Recursive Feature Elimination (RFE) algorithm with linear C-Support Vector Classification (SVC). The method was first adopted in bioinformatics and has now become one of the most common feature selection methods for machine learning tasks. SVC is a configurable classifier implemented based on libsvm (Chang & Lin, 2011) using one vs one scheme, which also suits the goal of the study on the three one vs one classification tasks. RFE finds a subset of features (top *X*) that contribute most to linear model training by recursively reducing the dimensionality of weight vector, where weight vector of feature coefficients is an attribute of linear SVM classifier.

We employed SVC-RRFE, which is a novel feature selection approach. Unlike the SVM-RFE diagram proposed in Guyon's work, the top *X* feature set was not selected from 689 altogether. Since the purpose is to investigate what linguistic features could help with identifying healthy control, MCI and SCI subjects. Despite the goal of finding the most optimal feature set to identify MCI and SCI individuals from healthy people, we would like to observe how each type of feature perform to the classification tasks. More specifically, how good could one type of features predict, and how much improvement could obtain by combining the feature groups. SVC-RRFE was employed on four feature groups respectively for each classification task, where all candidates were divided into morphological, syntactic, non-syllable related phonological features and other features (e.g. self-correction, unclear, repetition, interruption, MLU, MLW, ASL and ASC).

Instead of eliminating the feature subset from top to bottom, the SVC-RRFE method applies binary search to measure the performance of feature sets. The algorithm begins with training and predicting average accuracy and ROC value of group k-fold split dataset with multiple k values. It removes features that have 0 coefficient from the best k-fold model and keep the index and name of other features with the order of training contribution. The higher the coefficient, the higher rank it stays. The size of retained features becomes the biggest potential top *X* feature set in elimination. Then we apply binary search on it, keep the index of subsets and predict on them. Repeat the search until the superior feature set is found. The following equation applies to the whole feature elimination process:
Equation 1)

$$potential\ top\ X = \frac{size\ of\ top\ m\ +\ size\ of\ top\ n}{2} \tag{10}$$

The feature elimination procedure consists in eight (8) steps. It could be terminated at any step if all potential feature sets are examined.

Define:

1. Start set ***X***;
2. Group k-fold training and prediction method ***pred(feature,k)***, prints out average performance of each k fold;
3. Binary search method ***bi_search(m, n)***, returns outcome of equation;
4. Subset generating method ***top_x(x,feature_name,index_to_keep)***, where ***x*** stands for the size of subset, ***feature_name*** is the name of features in specific feature group before elimination, and ***index_to_keep*** is the index of set ***X*** features that is calculated by removing the index of 0 coefficient

feature after the first prediction with all the available features in defined feature group. It returns a subset with the feature name and their index from the original feature group.

Step 1. Generate three candidate feature set with ***bi_search(1,X)***, ***bi_search(bi_search(1,X),X)***, ***bi_search(1,bi_search(1,X))***, here we mark the result as ***A***, ***B*** and ***C***. Input ***A***, ***B*** and ***C*** respectively to ***top_x*** method to generate subset and use ***pred*** method with multiple ***k*** values. Take down the set with highest accuracy or roc.

Step 2. Mark the one that outperforms as ***Y***. If the one that outperforms is ***A***, we mark it as the best feature set and predict on **top_x(bi_search(A, B),feature_name,index_to_keep)** or to see if ***Y*** should be updated. If the one with the best performance is ***B or C***, check performance of **top_x(bi_search(C, B), feature_name, index_to_keep)**.

Step 3. If ***Y is not*** updated, call ***bi_search*** filling the current best feature size and the last experimented size to get the size of the next potential one. If ***Y*** updated, get the next set by calling ***bi_search*** with ***Y*** and the second-best set. Train and compare performance, take down the updated/remained ***Y*** and the performance of current feature set.

Step 4-5. Repeat step 3 twice to reduce the size of feature set and update ***Y***.

Step 6. Train and predict the closest 4 feature sets (**top *Y-2***, **top *Y-1***, **top *Y+1*** and **top *Y+2***) of updated ***Y*** from step 4, take down the set with best performance.

Step 7-8. If a better feature set gets discovered during step 5, repeat step 5 until no more improvement. Store the best set and feature elimination is finished.

The best features for classification tasks were then generated by the process described above on the combination of the best set from each feature group. With SVC-RRFE, the top *X* feature set could be concluded by only checking 10 different feature compositions out of more than 600 candidates.

Not all features were included during feature selection. Features that applied to the following criterion were eliminated:

- Features that are too rare. The “preposition error” was removed based on the principle since it only happened once in MCI individuals and HC and did not occur in SCI ones in resampled data.
- Features that did not get any hit except 0 in all samples after standardization. The five candidates that got removed are “g_uni_1”, “g_uni_2”, “uːp_bi_2_3”, “uːɔ_bi_2_3” and “yː_uni_1”.
- Features that gain 0 correlation coefficient value during feature elimination. The feature to be removed varies based on the elimination strategy (Parsey-Acc, Parsey-ROC, Sparv-Acc, Sparv-ROC). For example, 75 phonological features were eliminated by all elimination strategies which includes 70 biphones and 5 unigrams, where 17 of them only exist in MCI class, 37 ones only exist in SCI class and 21 ones only exist in HC class. All the other features (morphological, syntactic, dialog, syllabus) that got eliminated from the feature groups followed the same principle.

Model training and Model selection

Model training - k-Fold cross-validation

To evaluate the selected features from elimination process and identify the best models for the classification, the outcome was treated as input of feed forward neural networks. Similar to the ***pred*** method, the neural networks were trained with group k-fold that splitted and standardized data, where group in this context, is the code of the speaker that produced the data point. The structures of neural networks were implemented by module Keras (in python 3.6) which is a high-level API (Chollet, 2015) for Google's TensorFlow (Abadi et al., 2016). Most models were trained using 10-fold group cross-validation and in a few cases, we employed 3-fold group cross-validation to achieve better performance. All k models were saved and loaded during training to calculate the average performance of neural network configuration. The epoch size was set to 45 for all tasks, and batch size was assigned to 0.1 times of training set size. The optimization algorithm for all model compiling was Nesterov Adam optimizer (Nadam) (Dozat, 2016) with default learning rate of 0.002 and we used categorical cross entropy as the loss function.

For training efficiency, early stopping technique was adapted to finish training before the defined batch if no validation improvement occurs. It monitors validation accuracy for HC vs MCI subject classification, and validation loss to differentiate HC vs SCI subjects, and MCI vs SCI subjects. Metrics selection for monitoring were determined by the actual training process to gain better prediction results for specific classification tasks. The patience value of early stopping was set to 5 epochs. That is, the training process stops and saves the best model when the monitored metric shows no improvement after training for 5 times.

Model selection - Evaluation Metrics

The machine learning models were evaluated using the following evaluation metrics:

Matthew's Correlation Coefficients (MCC). MCC is a model evaluation measurement that has a variation of values between -1 and +1. When MCC is -1 it represents completely wrong classification between true class and actual prediction, and vice versa. When the value is 0, the prediction is considered random. The equation of MCC is:
Equation 1)

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (FN + TN) \times (FP + TN) \times (TP + FN)}} \tag{11}$$

Where TP = True Positive (the case that a person is sick was identified as sick), FN = False Negative (the case that a person is sick but was identified as healthy), FP = False Positive (the case that a person is healthy but was identified as sick), TN = True Negative (the case that a person is healthy and was identified as not sick). The reason to introduce MCC is because the calculation of accuracy includes True Negative and the result would not make sense if True Positive is 0. In addition, if False Negative is 0, recall could be 1.0 but the model might still predict by chance. However, when there is no FN and FP the prediction is guaranteed to be perfect and vice versa.

Area under Receiver operating characteristic (ROC AUC). ROC AUC is a performance measurement for classifiers. It measures how well predictions are ranked (proportion of positives should rank before negatives) and their quality. The ROC is a curve drawn by against True Positive rate (sensitivity) with False Positive rate (1-recall) and AUC is the area (shadow) under the curve. True Positive rate is calculated by

$\frac{TP}{TP+FN}$and False Positive rate is calculated by $\frac{FP}{FP+TN}$. The value varies from 0 to 1. The higher the value, the better the model is at classification and 0.5 means by chance.

The Area under ROC of the best models are demonstrated to evaluate the quality of them (see figure 7, 12, 17). The X-axis represents False positive rate and the Y-axis represents True positive rate of prediction. The True positive rate is also called sensitivity, which is calculated by dividing true positive predictions with all predictions that were positive. The False positive rate is known as 1-specificity, which comes from the output of false positive predictions divided by all negative ones. When identifying disease, high sensitivity means we have a higher chance to diagnose the person who has the disease and is sick, high specificity means we manage to identify the person who does not carry the disease is not sick.

When both the MCC and ROC AUC are high, we consider the model with high reliability.

Results

This chapter demonstrates the result from feature selections to neural network models for the three classification tasks.

Table 3-6, 9-12, 15-18 demonstrate the number of features and folds that were employed during feature selection in respect to the evaluation metrics described in the last chapter. We then observe the top 30 most important features for each model in figure 3-6, 8-11, 13-16.

For each selected feature group, we trained three neural network models and every neural network configuration was retrained three times to maximize the performance. Since the quantity of data is small for deep learning, the training output could never be exactly the same with the same structure. By retraining several times, the outcome becomes more reliable. We show the evaluation metrics of the neural network models trained by selected features in table 7, 13, 19, where the model name, mean and standard deviation of validation accuracy, mean and standard deviation of Matthews correlation coefficient and mean and standard deviation of area under the ROC (ROC_AUC) are presented together with the feature selection strategy. The best models are marked as bold in the table, and the area under ROC curves of the best models are showed in figure 7, 12, 17. In the end, we demonstrate the configuration of the best neural network models in table 8, 14, 20.

Study 1: HC vs MCI

Feature selection

Table 3 to 6 show the number of features and folds that were employed in each feature group corresponds to their validation MCC, validation ROC AUC and validation accuracy produced by the SVC classifier for HC vs MCI individual classification. The first 4 rows show the performance of single feature group (morphological, syntactic, phonological, dialog event and syllabus), and the last row shows the performance of the combined features. Benchmarks were generated based on the features after elimination and computed by the “metrics” class in scikit learn, which provides direct functions towards the validation MCC, validation ROC AUC and validation accuracy of the models.

Figure 3-6, 8-11, 13-16 show the top 30 features that contribute most to the classification tasks selected by different strategy (Parsey-Acc, Parsey-ROC, Sparv-Acc and Sparv-ROC). The X-axis stands for the correlation coefficients of the features printed from SVC classifier’s attribute “coef_” which indicates feature importance. The negative values and the positive values correspond to the negative class and positive class, respectively, and the representative changes based on the classification task. The Y-axis presents the name of the features. For phonological features, the name consists of the phoneme, the feature type (unigram or biphone) and its position in a token. For instance, ‘iːg_bi_2_3’ stands for words with [iːg] as biphone appears at the second and third position; ‘j_uni_2’ means words has [j] as unigram appears at the second position.

Text annotated using Parsey Universal, model selected by accuracy

Table 3: The number of features, folds and classification performance of feature groups using Parsey-Acc feature elimination strategy for HC vs MCI individual classification

Feature group	Number of features	Number of folds	MCC	ROC_AUC	Accuracy
Morphological features	9	4	-0.032	0.487	45.8%
Syntactic features	2	4	0	0.5	44%
Phonological features	255	4	0.281	0.642	63.7%
Other features (Dialog event + syllabus features)	3	8	0.173	0.575	58.8%
Morphological + Phonological + Other features	160	6	0.358	0.677	66.1%

The best model to identify HC vs MCI using Parsey-Acc approach was trained with group 6-fold cross validation. 160 features were selected to identify individuals with MCI from HC contains: count of self-correction, MLU, MLW, noun ratio, light verb ratio, adj ratio, infinitive ratio, 8 unigram positional segment frequencies and 145 bigram positional segment frequencies. By using only phonological features, we reached 63.7% accuracy and the accuracy was improved by 2% combining half of the morphological features, dialog features, MLU and MLW.

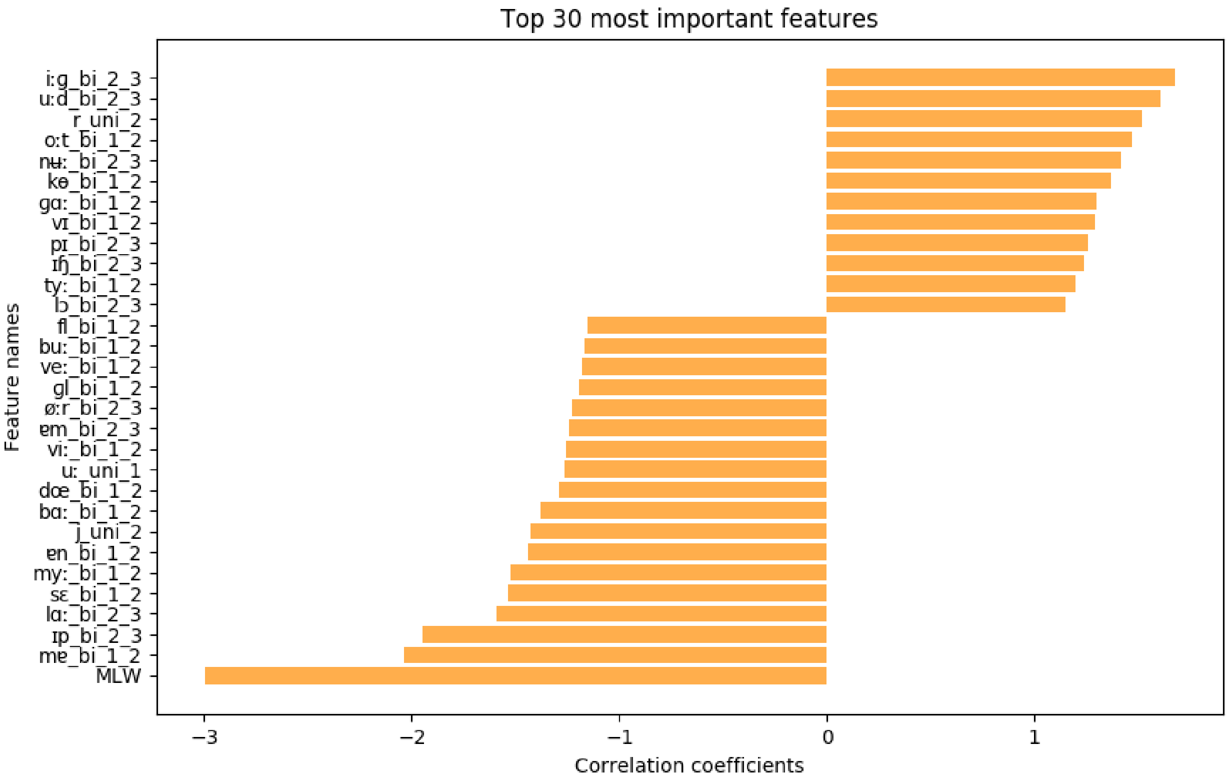

Figure 3: Top 30 features to classify HC and MCI group selected by Parsey-Acc. The x-axis shows the correlation coefficients of the feature and the y-axis shows the name of the feature. The longer the orange bar is, the higher contribution the feature makes. All bars to the left indicate negative classes and to the right positive classes.

Figure 3 shows the top 30 features that contribute most to classify HC and MCI group selected by Parsey-Acc strategy. The negative values indicate the negative class which in this case is HC, and the positive values contribute to the MCI class.

The top features for MCI group were: ‘lɔ_bi_2_3’, ‘tyː_bi_1_2’, ‘ɪɧ_bi_2_3’, ‘pɪ_bi_2_3’, ‘vɪ_bi_1_2’, ‘gɑː_bi_1_2’, ‘kɵ_bi_1_2’, ‘nʉː_bi_2_3’, ‘oːt_bi_1_2’, ‘r_uni_2’, ‘uːd_bi_2_3’, ‘iːg_bi_2_3’. Words with [iːg] at the second and third position received the highest correlation coefficient.

The top features for HC group were: ‘MLW’, ‘mɐ_bi_1_2’, ‘ɪp_bi_2_3’, ‘lɑː_bi_2_3’, ‘sɛ_bi_1_2’, ‘myː_bi_1_2’, ‘ɐn_bi_1_2’, ‘j_uni_2’, ‘bɑː_bi_1_2’, ‘dœ_bi_1_2’, ‘uː_uni_1’, ‘viː_bi_1_2’, ‘ɐm_bi_2_3’, ‘øːr_bi_2_3’, ‘gl_bi_1_2’, ‘veː_bi_1_2’, ‘buː_bi_1_2’, ‘fl_bi_1_2’. MLW is the strongest feature to identify HC from individual with MCI.

Most features in the top 30 feature set were phonological ones, where 42% features for MCI subjects were biphones that appear at second and third index of words, such as [uːd], [nʉː], [pɪ], [ɪɧ] and [lɔ], etc. In addition, the ‘pɪ_bi_2_3’ feature was only found in MCI individual samples which made it a strong indicator to identify MCI subjects from HC and SCI. Another significant predictor is ‘ɪp_bi_2_3’ which only existin HC sample and was one of the top features for HC.

Text annotated using Parsey Universal, model selected by area under ROC and Matthews correlation coefficient

Table 4: The number of features, folds and classification performance of feature groups using Parsey-ROC feature elimination strategy for HC vs MCI individual classification

Feature group	Number of features	Number of folds	MCC	ROC_AUC	Accuracy
Morphological features	4	10	0.057	0.52	43.3%
Syntactic features	2	10	0.009	0.503	38.9%
Phonological features	252	10	0.29	0.65	63.8%
Other features (Dialog event + syllabus features)	3	8	0.173	0.575	58.8%
Morphological + Phonological + Other features	216	10	0.358	0.684	66%

The 216 most important features to identify individuals with MCI from HC using Parsey-ROC approach were selected by group 10-fold cross validation, including count of MLW, ratio of infinitives, 10 unigram positional segment frequencies and 204 bigram positional segment frequencies.

With only phonological predictors, the SVM classifier identified MCI subjects with 63.7% accuracy, which contribute the most to the feature selection with combined feature groups. While the final number of features for classification were different (160 and 216), the accuracy and MCC were almost the same (66.1%

and 66%; 0.358 and 0.358). However, the ROC AUC score indicated that the second model performed better.

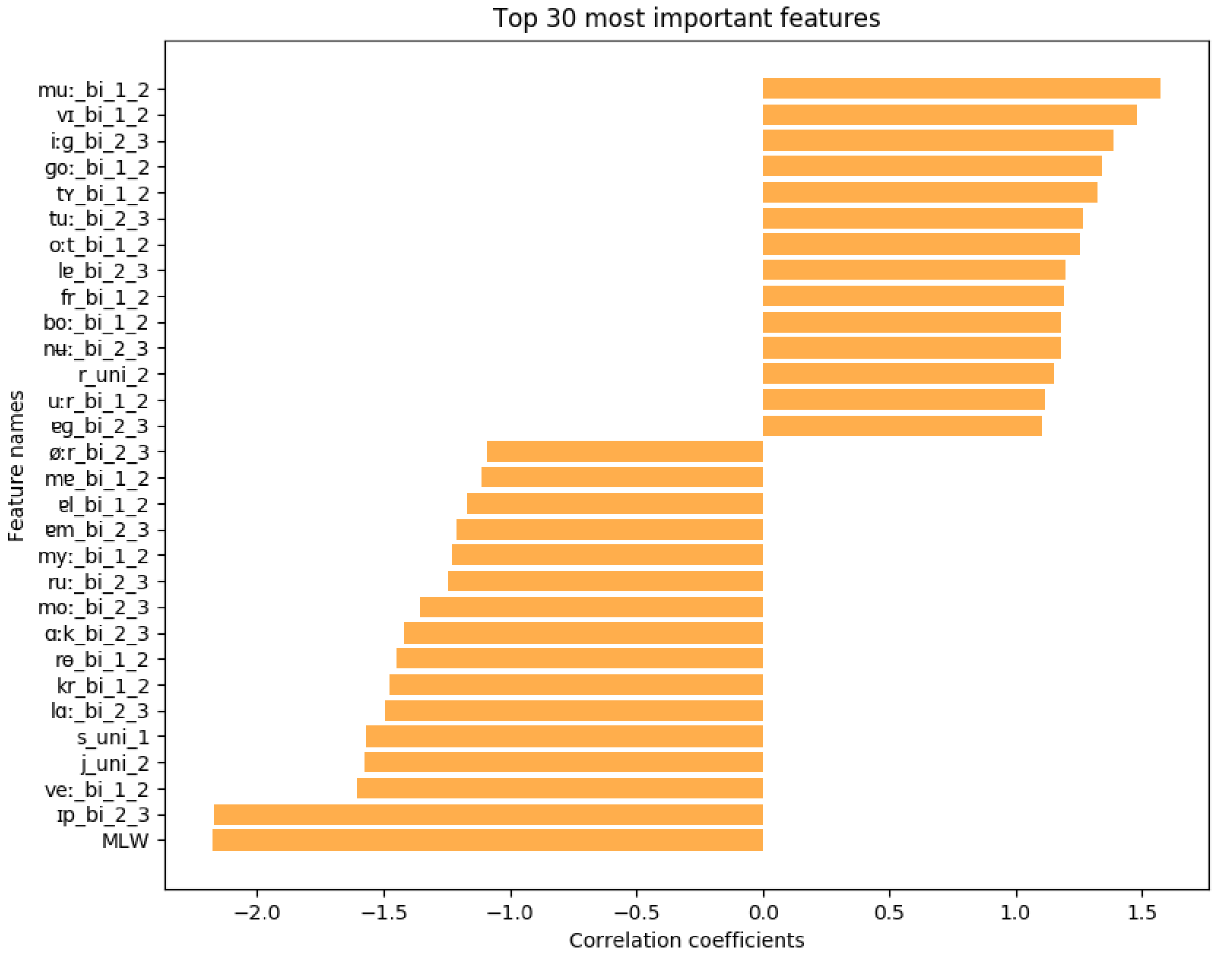

Figure 4: Top 30 features to classify HC and MCI group selected by Parsey-ROC. The x-axis shows the correlation coefficients of the feature and the y-axis shows the name of the feature. The longer the orange bar is, the higher contribution the feature makes. All bars to the left indicate negative classes and to the right positive classes.

Figure 4 shows the top 30 features that contribute most to classify HC and MCI group selected by Parsey-ROC strategy. The negative values indicate the negative class which in this case is HC, and the positive values contribute to the MCI class. The top features for MCI group were: ‘ɐg_bi_2_3’, ‘uːr_bi_1_2’, ‘r_uni_2’, ‘nʉː_bi_2_3’, ‘boː_bi_1_2’, ‘fr_bi_1_2’, ‘lɐ_bi_2_3’, ‘oːt_bi_1_2’, ‘tuː_bi_2_3’, ‘tʏ_bi_1_2’, ‘goː_bi_1_2’, ‘iːg_bi_2_3’, ‘vɪ_bi_1_2’, ‘muː_bi_1_2’. Words begin with [muː] was the most important feature to detect individuals with MCI.

The top features for HC group were: ‘MLW’, ‘ɪp_bi_2_3’, ‘veː_bi_1_2’, ‘j_uni_2’, ‘s_uni_1’, ‘lɑː_bi_2_3’, ‘kr_bi_1_2’, ‘rɵ_bi_1_2’, ‘ɑːk_bi_2_3’, ‘moː_bi_2_3’, ‘ruː_bi_2_3’, ‘myː_bi_1_2’, ‘ɐm_bi_2_3’, ‘ɐl_bi_1_2’, ‘mɐ_bi_1_2’, ‘øːr_bi_2_3’. Same as the output from Parsey-Acc approach, the MLW feature is the top predictor for HC.

The proportion of MCI: HC in top 30 features was 0.875. For MCI group, 11 predictors were biphones and 1 was unigram PSF. The biphone PSF feature ‘uːr_bi_1_2’ only exist in MCI samples and ‘ɪp_bi_2_3’ only exist in HC samples which made them one of the strongest predictors for MCI subjects and HC, respectively.

Comparing with the model trained with Parsey-Acc method, features such as ‘MLW’, ‘j_uni_2’, ‘lɑː_bi_2_3’, ‘myː_bi_1_2’, ‘mɐ_bi_1_2’, ‘veː_bi_1_2’, ‘øːr_bi_2_3’, ‘ɐm_bi_2_3’ and ‘ɪp_bi_2_3’ are the common top predictors for HC and ‘iːg_bi_2_3’, ‘nʉː_bi_2_3’, ‘oːt_bi_1_2’, ‘r_uni_2’, ‘vɪ_bi_1_2’ are the common top ones for MCI.

Text annotated using Sparv, model selected by accuracy

Table 5: The number of features, folds and classification performance of feature groups using Sparv-Acc feature elimination strategy for HC vs MCI individual classification

Feature group	Number of features	Number of folds	MCC	ROC_AUC	Accuracy
Morphological features	4	4	0.02	0.505	45.6%
Syntactic features	2	4	-0.027	0.5	43.8%
Phonological features	313	10	0.292	0.652	64.6%
Other features (Dialog event + syllabus features)	3	8	0.173	0.575	58.8%
Morphological + Phonological + Other features	235	10	0.403	0.71	68.7%

Same as the Parsey-ROC approach, the best features to identify individuals with MCI from HC by Sparv-Acc were selected by group 10-fold cross validation. MLW, infinitive ratio, 14 unigram positional segment frequencies and 219 bigram positional segment frequencies consisted of 235 most important features.

With only phonological predictors, the model identified MCI subjects with 64.6% accuracy. Although 80 of them were eliminated, the find accuracy was improved by 4% by combining the rest of the phonological features with one morphological feature and MLW. The MCC was increased from 0.292 to 0.403 which made the last model more reliable.

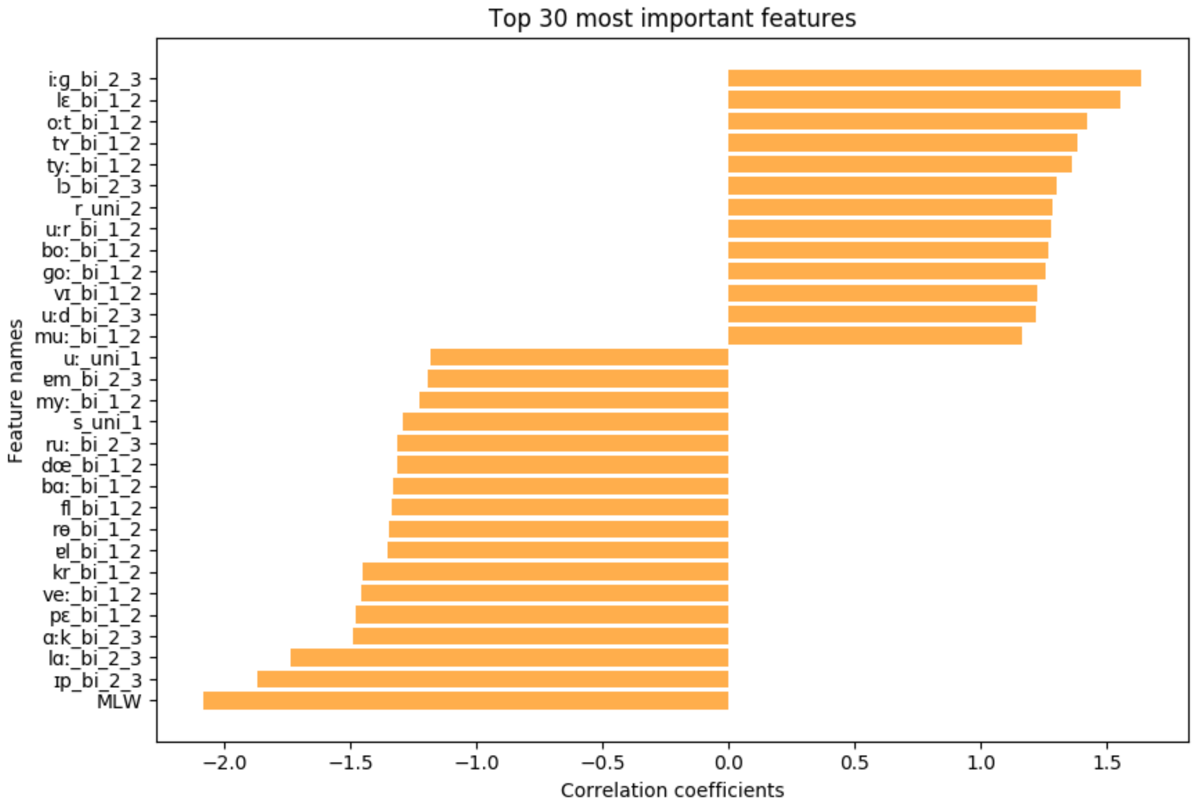

Figure 5: Top 30 features to classify HC and MCI group selected by Sparv-Acc. The x-axis shows the correlation coefficients of the feature and the y-axis shows the name of the feature. The longer the orange bar is, the higher contribution the feature makes. All bars to the left indicate negative classes and to the right positive classes.

Figure 5 shows the top 30 features that contribute most to classify HC and MCI group selected by Sparv-Acc strategy. The negative values indicate the negative class which in this case is HC, and the positive values contribute to the MCI class.

The top features for MCI group were: ‘tuː_bi_2_3’, ‘kɔ_bi_1_2’, ‘lɔ_bi_2_3’, ‘boː_bi_1_2’, ‘vɪ_bi_1_2’, ‘lɐ_bi_2_3’, ‘oːt_bi_1_2’, ‘r_uni_2’, ‘muː_bi_1_2’, ‘goː_bi_1_2’, ‘lɛ_bi_1_2’ and ‘iːg_bi_2_3’.

The top features for HC group were: MLW, ‘ɪp_bi_2_3’, ‘veː_bi_1_2’, ‘kr_bi_1_2’, ‘pɛ_bi_1_2’, ‘lɑː_bi_2_3’, ‘bɑː_bi_1_2’, ‘dœ_bi_1_2’, ‘s_uni_1’, ‘rɵ_bi_1_2’, ‘moː_bi_2_3’, ‘ɐm_bi_2_3’, ‘ɑːk_bi_2_3’, ‘ɐl_bi_1_2’, ‘ruː_bi_2_3’, ‘myː_bi_1_2’, ‘mt_bi_2_3’ and ‘bɛː_bi_1_2’.

Same as the features selected by Parsey-Acc approach, words with [iːg] at the second and third position was the most important predictor for MCI subjects and MLW is the strongest feature for HC. Among the top 30 best features, 29 were introduced by the phonological feature group and the proportion of MCI: HC was 0.67. Comparing all features selected by accuracy, MLW together with ‘lɑː_bi_2_3’, ‘ɐm_bi_2_3’, ‘ɪp_bi_2_3’, ‘veː _bi_1_2’, ‘myː_bi_1_2’ were the common predictors for HC; ‘r_uni_2’, ‘oːt_bi_1_2’, ‘vɪ_bi_1_2’ and ‘iːg_bi_2_3’ were the common features for MCI group.

Text annotated using Sparv, model selected by area under ROC and Matthews correlation coefficient

Table 6: The number of features, folds and classification performance of feature groups using Sparv-ROC feature elimination strategy for HC vs MCI individual classification

Feature group	Number of features	Number of folds	MCC	ROC_AUC	Accuracy
Morphological features	6	8	0.069	0.533	47.2%
Syntactic features	2	10	0.006	0.501	38.3%
Phonological features	252	10	0.290	0.650	63.8%
Other features (Dialog event + syllabus features)	3	8	0.173	0.575	58.8%
Morphological + Phonological + Other features	201	10	0.353	0.683	66%

The best features to identify individuals with MCI from HC were selected by group 10-fold cross validation, same as the Parsey-ROC and Sparv-Acc approach. The 201 features included MLW, noun ratio, infinitive ratio, 8 unigram positional segment frequencies and 190 bigram positional segment frequencies.

With only phonological predictors, the model identified MCI subjects with 63.8% accuracy (slightly lower than Sparv-Acc method). The overall performance was 66% accuracy with 0.353 MCC and 0.683 ROC AUC.

When it comes to model performance for text annotated by Sparv, the model selected with accuracy overall performed better than the one selected with Area under ROC. Although the MCC of morphological feature group in the later strategy (0.069) was higher than accuracy-oriented one (0.02) and the MCCs of other event feature group were very close in both cases, model selection with accuracy got better result (Sparv-ROC: 66% accuracy, 0.683 ROC_AUC, 0.353 MCC, Sparv-Acc: 68.7% accuracy, 0.71 ROC_AUC, 0.403 MCC). It turns out that a better construction of a phonological feature group is the key to a better model and the phonological feature group played the most important role in all feature selection tasks.

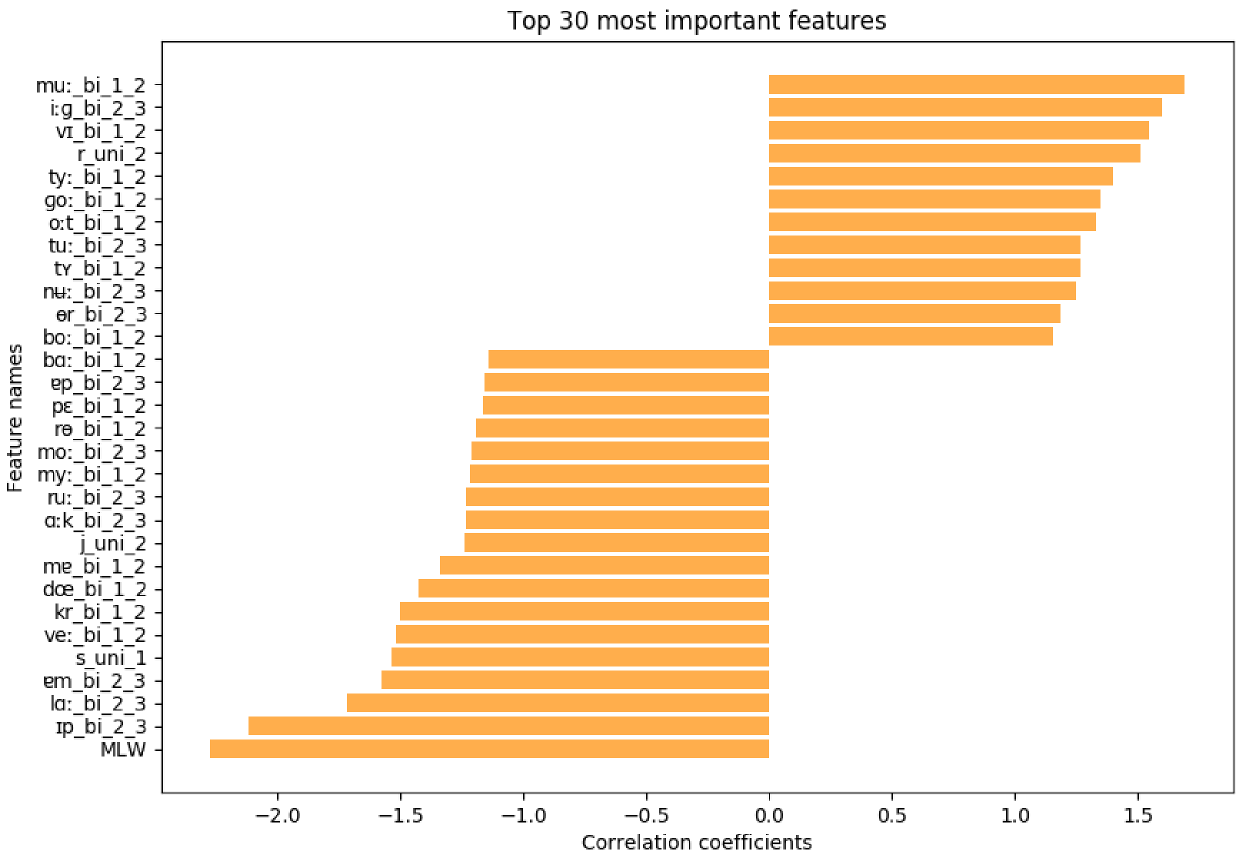

Figure 6: Top 30 features to classify HC and MCI group selected by Sparv-ROC. The x-axis shows the correlation coefficients of the feature and the y-axis shows the name of the feature. The longer the orange bar is, the higher contribution the feature makes. All bars to the left indicate negative classes and to the right positive classes.

Figure 6 shows the top 30 features that contribute most to classify HC and MCI group selected by Sparv-ROC strategy. The negative values indicate the negative class which in this case is HC, and the positive values contribute to the MCI class.

Same as all the other feature selection approaches, MLW was the most import feature and one of the common features for HC, where the other common top features were ‘lɑː_bi_2_3’, ‘myː_bi_1_2’, ‘veː_bi_1_2’, ‘ɐm_bi_2_3’ and ‘ɪp_bi_2_3’. It is worth noting that ‘ɪp_bi_2_3’ as a feature only exist in HC samples was included in all models as top HC predictor. The common top features from all feature selection approaches for MCI group are ‘iːg_bi_2_3’, ‘oːt_bi_1_2’, ‘r_uni_2’ and ‘vɪ_bi_1_2’. Though the order of the importance differs, biphone PSF feature ‘iːg_bi_2_3’ always stay in the top 3 to identify MCI individual from HC. The number of MCI features in the top 30 feature list are very close where both Parsey-Acc and Sparv-ROC have 12 one, Sparv-Acc has 13 ones and Parsey-ROC has 14 ones. 10 top features were the same between Sparv-ROC and Parsey-ROC and 9 were identical in Sparv-ROC and Sparv-Acc with slightly different rankings of importance. Most of the common features are the same except ‘nʉː_bi_2_3’ was not on the top 30 list in Sparv-Acc model.

Regardless of the annotation tool, using ROC_AUC as a benchmark for model selection resulted in similar performance for MCI individual detection in contrast to using accuracy as benchmark.

Feed forward neural networks

The following chart shows the statistics of 12 NN models built upon 4 selected feature groups for HC vs MCI classification:

Table 7: The validation statistics of NN models in respect to the feature elimination strategy for HC vs MCI individual classification

Model	Validation accuracy mean	SD	Matthews correlation coefficient mean	ROC AUC mean	SD	Feature selection measurement	Annotation tool
M1	74.34	6.39	0.45	0.73	0.08	ROC AUC	Parsey
M2	72.01	6.74	0.41	0.71	0.08	ROC AUC	Parsey
M3	70.86	6.65	0.40	0.71	0.05	ROC AUC	Parsey
M4	73.45	4.84	0.42	0.71	0.07	ROC AUC	Sparv
M5	71.51	4.26	0.41	0.71	0.07	ROC AUC	Sparv
M6	71.85	6.24	0.41	0.71	0.08	ROC AUC	Sparv
M7	**75.84**	**6.69**	**0.47**	**0.73**	**0.08**	**Accuracy**	**Parsey**
M8	73.90	4.41	0.44	0.72	0.07	Accuracy	Parsey
M9	75.16	5.95	0.45	0.72	0.08	Accuracy	Parsey
M10	72.54	5.64	0.39	0.7	0.10	Accuracy	Sparv
M11	72.78	6.09	0.44	0.73	0.07	Accuracy	Sparv
M12	72.52	4.74	0.42	0.71	0.07	Accuracy	Sparv

Figure 7 demonstrates the area under Receiver Operating Characteristic (ROC) curve of model M7:

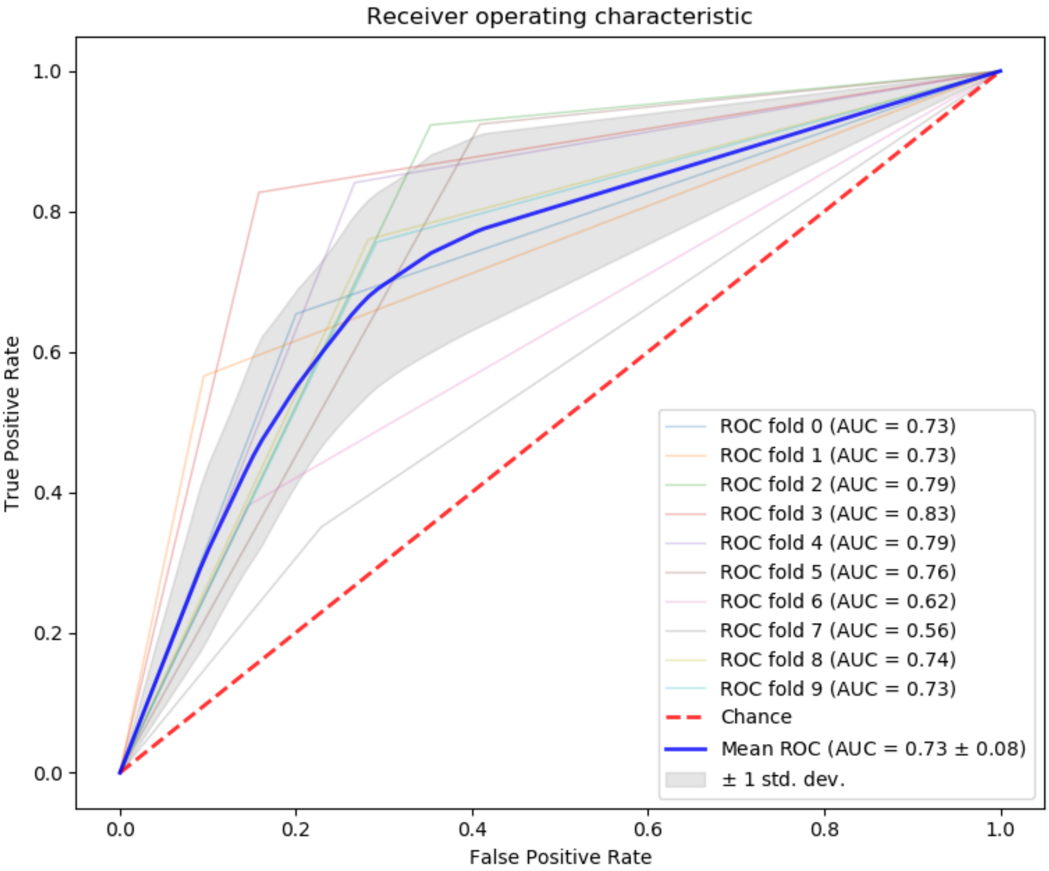

Figure 7: Mean ROC curve and AUC of the 10-fold cross validation model M7 for HC vs MCI classification. The x-axis shows the false positive rate and the y-axis shows the true positive rate of the evaluation. The 10 light color curves correspond to the ROCs for each one of the 10 folds. The bold blue curve shows the mean ROC and the grey shadow shows the standard deviation of the mean ROC curve. The red dotted line represents the baseline that predicts the class by chance. Curves equal or below the red dotted line indicate a bad model.

The best model is M7 using 160 features selected from Parsey Universal annotated text with accuracy. It resulted into 75.84% (+/- 6.69%) mean validation accuracy with 0.73 (+/- 0.08) mean ROC AUC and 0.47 mean Matthews correlation coefficient score. The improvement of performance is significant compared to the model from feature selection, where the accuracy increased by 12.8%, the ROC AUC increased by 7.3% and the MCC increased by 23.8%. Although the highest accuracy occurred at fold 6 which reached 84.1%, the best output among 10 attempts was fold 3 with 0.83 ROC AUC score and 81.4% accuracy. The worst model was the one trained using as a 7-fold cross-validation while it presents a slightly better performance in comparison to the baseline (0.5) with 63% accuracy and 0.56 ROC AUC.

Table 8 presents the configuration of the M7 model, which contains one input layer, one hidden layer and one output layer. The input layer scales 160 input dimensions (number of feature columns generated by Parsey-Acc approach to identify MCI individuals from HC during feature selection) to 350 fully connected neurons with hyperbolic tangent (tanh) as activation function and 0.6 dropout. The hidden layer has 30 neurons and its activation function is Rectified Linear units (ReLu). A dropout layer was added to set 20% of the activations to zero before the output layer. In the end, we applied the Softmax function on the output layer to calculate the probabilities of classification result.

Table 8: The configuration of the M7 model for HC vs MCI individual classification

Layer	Dimension	Activation	Dropout
Input layer	350 (input 160 dimensions)	tanh	0.6
Hidden layer	30	ReLu	0.2
Output layer	2	Softmax	N/A

Study 2: HC vs SCI

Feature selection

Table 9-12 demonstrates the number of features and folds that were employed in each feature group corresponds to their validation MCC, validation ROC AUC and validation accuracy produced by the SVC classifier for HC vs SCI individual classification. The composition of the table is the same as the ones in Study 1.

Text annotated using Parsey Universal, model selected by accuracy

Table 9: The number of features, folds and classification performance of feature groups using Parsey-Acc feature elimination strategy for HC vs SCI individual classification

Feature group	Number of features	Number of folds	MCC	ROC_AUC	Accuracy
Morphological features	2	3	0.057	0.51	55.1%
Syntactic features	2	3	0.021	0.511	51.5%
Phonological features	234	12	0.300	N/A	68%
Other features (Dialog event + syllabus features)	3	12	0.075	N/A	55.1%
Phonological + Other features	150	12	0.294	N/A	68.8%

Since the distribution of the validation class in this case has only HC labels (marked as 0) using 12 folds cross validation, true positive value became meaningless and the ROC_AUC could not be computed and presented in phonological feature group and other feature group.

The best feature group combination selected by Parsey-Acc approach to distinguish SCI individuals from HC included phonological features, count of self-correction and MLU. The average MLU of SCI group (14.85) is higher than HC group (13.05) which makes it a comparable strong predictor to detect SCI individuals. By using only phonological features, we reached 68% accuracy and the accuracy was only improved by 0.8% by adding MLU and self-correction.

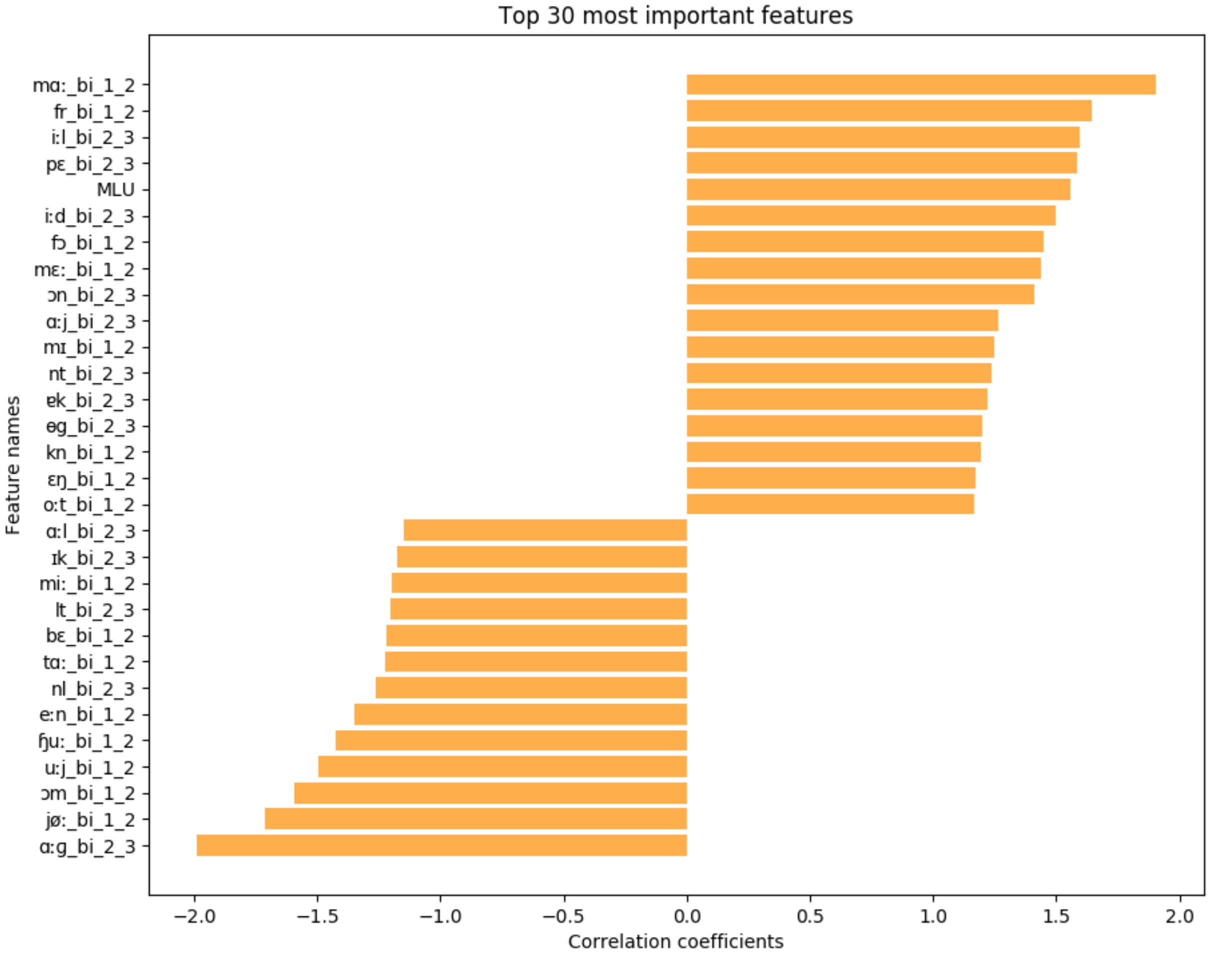

Figure 8: Top 30 features to classify HC and SCI group selected by Parsey-Acc. The x-axis shows the correlation coefficients of the feature and the y-axis shows the name of the feature. The longer the orange bar is, the higher contribution the feature makes. All bars to the left indicate negative classes and to the right positive classes.

Figure 8 shows the top 30 features that contribute most to classify HC and SCI group selected by Parsey-Acc strategy. The negative values indicate the negative class which in this case is HC, and the positive values contribute to the SCI class. The top predictors for SCI subjects were: ‘oːt_bi_1_2’, ‘ɛŋ_bi_1_2’, ‘kn_bi_1_2’, ‘ɵg_bi_2_3’, ‘ɐk_bi_2_3’, ‘nt_bi_2_3’, ‘mɪ_bi_1_2’, ‘ɑːj_bi_2_3’, ‘ɔn_bi_2_3’, ‘mɛː_bi_1_2’, ‘fɔ_bi_1_2’, ‘iːd_bi_2_3’, ‘MLU’, ‘pɛ_bi_2_3’, ‘iːl_bi_2_3’, ‘fr_bi_1_2’, ‘mɑː_bi_1_2’.

The top features to predict HC were: ‘ɑːg_bi_2_3’, ‘jøː_bi_1_2’, ‘ɔm_bi_1_2’, ‘uːj_bi_1_2’, ‘ɧuː_bi_1_2’, ‘eːn_bi_1_2’, ‘nl_bi_2_3’, ‘tɑː_bi_1_2’, ‘bɛ_bi_1_2’, ‘lt_bi_2_3’, ‘miː_bi_1_2’, ‘ɪk_bi_2_3’, ‘ɑːl_bi_2_3’

Among the top 30 most important features, 29 are phonological predictors where all came of them are bigrams. Words starting with [mɑː] was the strongest predictor for SCI and words having [ɑːg] at the second and third index occurred more in the HC group. ‘uːj_bi_1_2’ could only be found in HC samples which is also the fourth strongest feature for HC.

Text annotated using Parsey Universal, model selected by area under ROC and Matthews correlation coefficient

Table 10: The number of features, folds and classification performance of feature groups using Parsey-ROC feature elimination strategy for HC vs SCI individual classification

Feature group	Number of features	Number of folds	MCC	ROC_AUC	Accuracy
Morphological features	6	8	0.051	0.517	54%
Syntactic features	2	6	0.046	0.526	50.4%
Phonological features	231	10	0.344	0.684	67.3%
Other features (Dialog event + syllabus features)	5	10	0.156	0.578	52.4%
Morphological + Phonological	175	10	0.373	0.701	69.1%

The most important features to identify individuals with SCI from HC with Parsey-ROC approach were selected by group 10-fold cross validation, which contained the ratio of infinitives, ratio of verbs, ratio of participle, ratio of adjectives, 9 unigram PSFs and 162 bigram PSFs.

The model identified SCI subjects with 54% accuracy by only morphological features and 67.3% accuracy by only phonological features. The overall performance was 69.1% accuracy (higher than Parsey-Acc approach) with 0.373 MCC and 0.701 ROC AUC with the combination of morphological and phonological predictors.

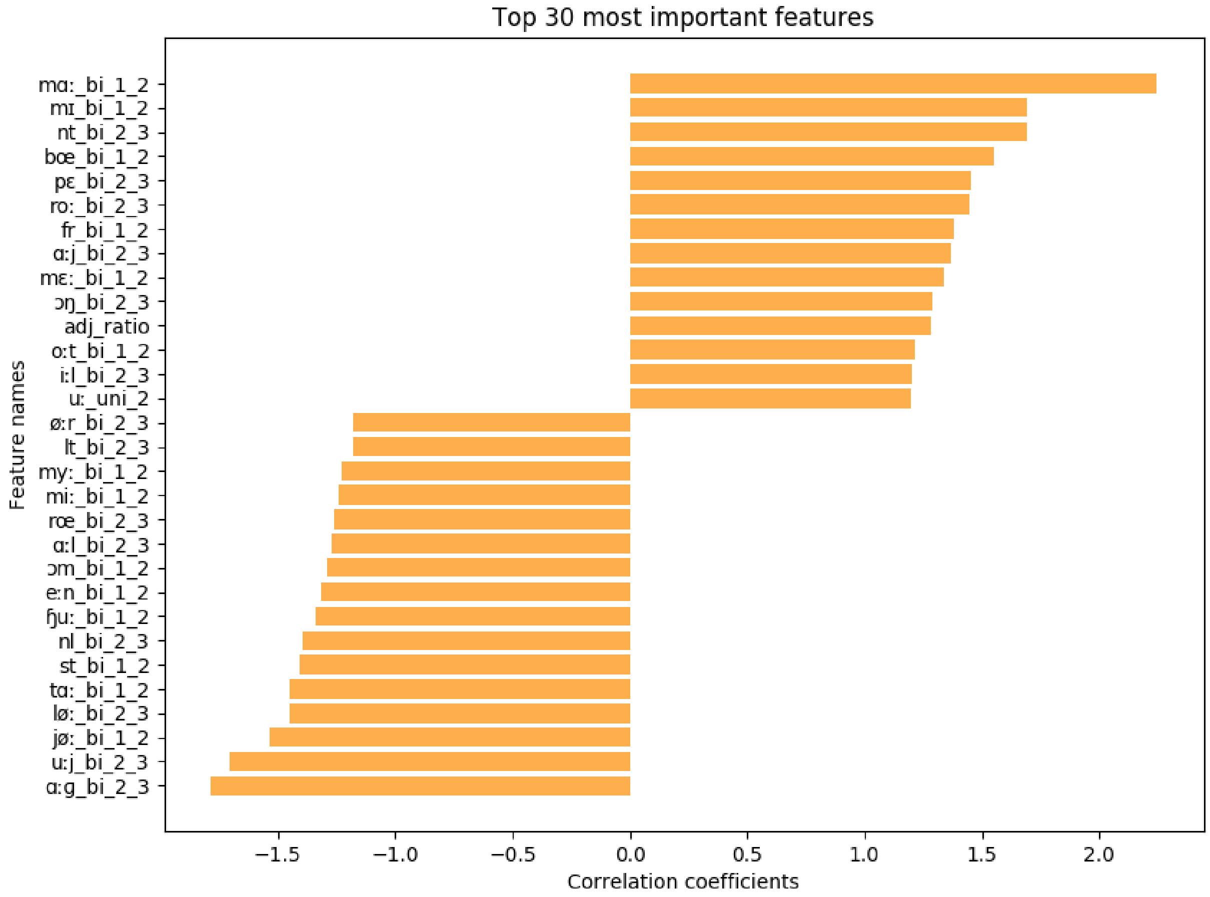

Figure 9: Top 30 features to classify HC and SCI group selected by Parsey-ROC. The x-axis shows the correlation coefficients of the feature and the y-axis shows the name of the feature. The longer the orange bar is, the higher contribution the feature makes. All bars to the left indicate negative classes and to the right positive classes.

Figure 9 shows the top 30 features that contribute most to classify HC and SCI group selected by Parsey-ROC strategy. The negative values indicate the negative class which in this case is HC, and the positive values contribute to the SCI class.
The top SCI features were: ‘uː_uni_2’, ‘iːl_bi_2_3’, ‘oːt_bi_1_2’, ‘ɔŋ_bi_2_3’, ‘mɛː_bi_1_2’, ‘ɑːj_bi_2_3’, ‘fr_bi_1_2’, ‘roː_bi_2_3’, ‘pɛ_bi_2_3’, ‘bœ_bi_1_2’, ‘nt_bi_2_3’, ‘mɪ_bi_1_2’ and ‘mɑː_bi_1_2’.

The top HC features were: ‘ɑːg_bi_2_3’, ‘uːj_bi_2_3’, ‘jøː_bi_1_2’, ‘løː_bi_2_3’, ‘tɑː_bi_1_2’, ‘st_bi_1_2’, ‘nl_bi_2_3’, ‘ɧuː_bi_1_2’, ‘eːn_bi_1_2’, ‘ɔm_bi_1_2’, ‘ɑːl_bi_2_3’, ‘rœ_bi_2_3’, ‘miː_bi_1_2’, ‘myː_bi_1_2’, ‘lt_bi_2_3’, ‘øːr_bi_2_3’.

In the top 30 most important features, 93% of predictors for SCI individuals came from phonological feature group, including. SCI individuals were observed using more adjectives than HC. Same as Parsey-Acc approach, the strongest predictor for SCI individuals was words begin with [mɑː] and words with [ɑːg] in the middle contributed most to HC.

Compared with the top 30 features of text annotated by Parsey Universal but features selected by accuracy, words begin with [fr], [mɑː], [mɛː], [mɪ], [oːt] and words with [iːl], [nt], [pɛ], [ɑːj] in the middle were the intersection of two feature groups for SCI individual detection; words begin with [eːn], [jøː], [miː], [tɑː], [ɔm], [ɧuː] and with [lt], [nl], [ɑːl], [ɑːg] at second and third index were common features for HC.

Text annotated using Sparv, model selected by accuracy

Table 11: The number of features, folds and classification performance of feature groups using Sparv-Acc feature elimination strategy for HC vs SCI individual classification

Feature group	Number of features	Number of folds	MCC	ROC_AUC	Accuracy
Morphological features	2	3	0.055	0.515	54.76%
Syntactic features	2	3	0.015	0.504	54.37%
Phonological features	236	6	0.342	0.67	67.32%
Other features (Dialog event + syllabus features)	2	5	0.130	0.572	54%
Morphological + Phonological	194	10	0.360	0.691	68.31%

Same as the Parsey-ROC approach, the best features selected by Sparv-Acc method to identify individuals with SCI from HC were selected by group 10-fold cross validation. Although 19 more features were included, the overall model performance was slightly worse than Parsey-ROC selected model. Ratio of adjectives, ratio of adverbs, 10 unigram PSFs and 182 bigram PSFs consisted of the best feature group. All morphological features were included, and both contributed to SCI group.

With only phonological predictors, the model identified SCI subjects with 67.32% accuracy. However, the combined feature model's performance was not much improved by adding morphological features and MLU. The MCC was increased to 0.360 and the ROC AUC was slightly increased to 0.691.

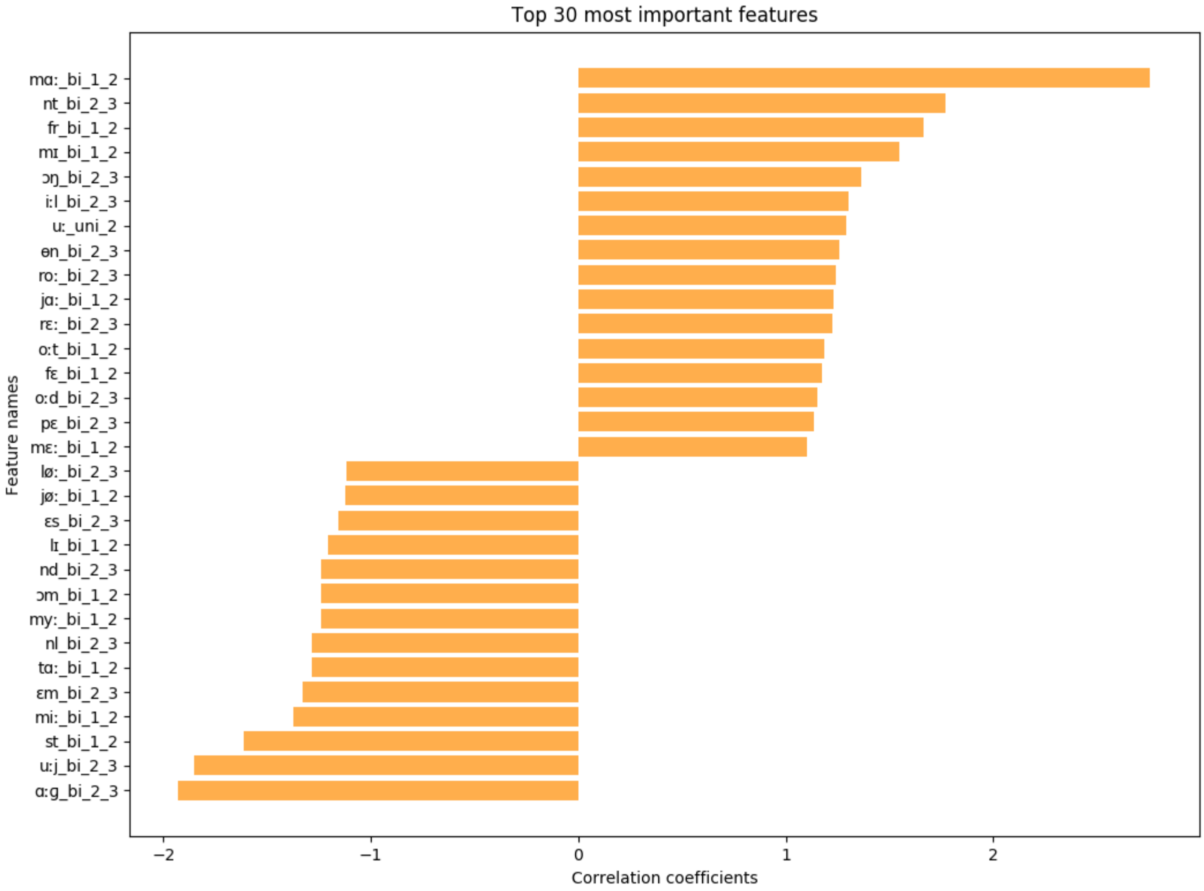

Figure 10: Top 30 features to classify HC and SCI group selected by Sparv-Acc. The x-axis shows the correlation coefficients of the feature and the y-axis shows the name of the feature. The longer the orange bar is, the higher contribution the feature makes. All bars to the left indicate negative classes and to the right positive classes.

Figure 10 shows the top 30 features that contribute most to classify HC and SCI group selected by Sparv-Acc strategy. The negative values indicate the negative class which in this case is HC, and the positive values contribute to the SCI class.

The top features for SCI group were: ‘mɛː_bi_1_2’, ‘pɛ_bi_2_3’, ‘oːd_bi_2_3’, ‘fɛ_bi_1_2’, ‘oːt_bi_1_2’, ‘rɛː_bi_2_3’, ‘jɑː_bi_1_2’, ‘roː_bi_2_3’, ‘ɵn_bi_2_3’, ‘uː_uni_2’, ‘iːl_bi_2_3’, ‘ɔŋ_bi_2_3’, ‘mɪ_bi_1_2’, ‘fr_bi_1_2’, ‘nt_bi_2_3’ and ‘mɑː_bi_1_2’.

The top features for HC group were: ‘ɑːg_bi_2_3’, ‘uːj_bi_2_3’, ‘st_bi_1_2’, ‘miː_bi_1_2’, ‘ɛm_bi_2_3’, ‘tɑː_bi_1_2’, ‘nl_bi_2_3’, ‘myː_bi_1_2’, ‘ɔm_bi_1_2’, ‘nd_bi_2_3’, ‘lɪ_bi_1_2’, ‘ɛs_bi_2_3’, ‘jøː_bi_1_2’ and ‘løː_bi_2_3’.

Among the top 30 most important features, ‘mɑː_bi_1_2’ and ‘ɑːg_bi_2_3’ were the strongest indicator for SCI and HC, respectively. Compared to the top 30 features selected from Parsey Universal annotated text with the same measurement, ‘fr_bi_1_2’, ‘mɑː_bi_1_2’, ‘mɛː_bi_1_2’, ‘mɪ_bi_1_2’, ‘oːt_bi_1_2’, ‘nt_bi_2_3’, ‘iːl_bi_2_3’ and ‘pɛ_bi_2_3’ are the common features for SCI individual and ‘jøː_bi_1_2’, ‘miː_bi_1_2’, ‘tɑː_bi_1_2’, ‘ɔm_bi_1_2’, ‘ɑːg_bi_2_3’and ‘nl_bi_2_3’ are the common ones for HC.

Text annotated using Sparv, model selected by area under ROC and Matthews correlation coefficient

Table 12: The number of features, folds and classification performance of feature groups using Sparv-ROC feature elimination strategy for HC vs SCI individual classification

Feature group	Number of features	Number of folds	MCC	ROC_AUC	Accuracy
Morphological features	5	8	0.088	0.530	54%
Syntactic features	2	6	0.015	0.504	54.4%
Phonological features	231	10	0.344	0.684	67.3%
Other features (Dialog event + syllabus features)	4	10	0.156	0.578	52.4%
Morphological + Phonological	175	10	0.370	0.7	68.8%

The best features generated by Sparv-ROC method to identify individuals with SI from HC were selected by group 10-fold cross validation, same as the Parsey-ROC and Sparv-Acc approach. The 175 features included 172 bigram phonological features and 3 phonological features. Ratio of adjectives and adverbs were indicators of SCI individuals and ratio of infinitives was predictor for HC.

The model identified SCI subjects with 54% accuracy by only morphological features and 67.3% accuracy by only phonological features, which is similar as models trained by Parsey-ROC and Sparv-Acc approach. The overall performance was 68.8% accuracy with 0.370 MCC and 0.7 ROC AUC.

Compared to model performance for text annotated by Sparv, the current model performed better than the one selected with accuracy. The overall best model was trained by Parsey-ROC approach with the highest accuracy, MCC and ROC AUC. Phonological features were always the strongest indicators (accuracy: 68%, 67.3%, 67.3%, 67.3%) for both classes which shows the confidence to distinguish SCI individuals from healthy people.

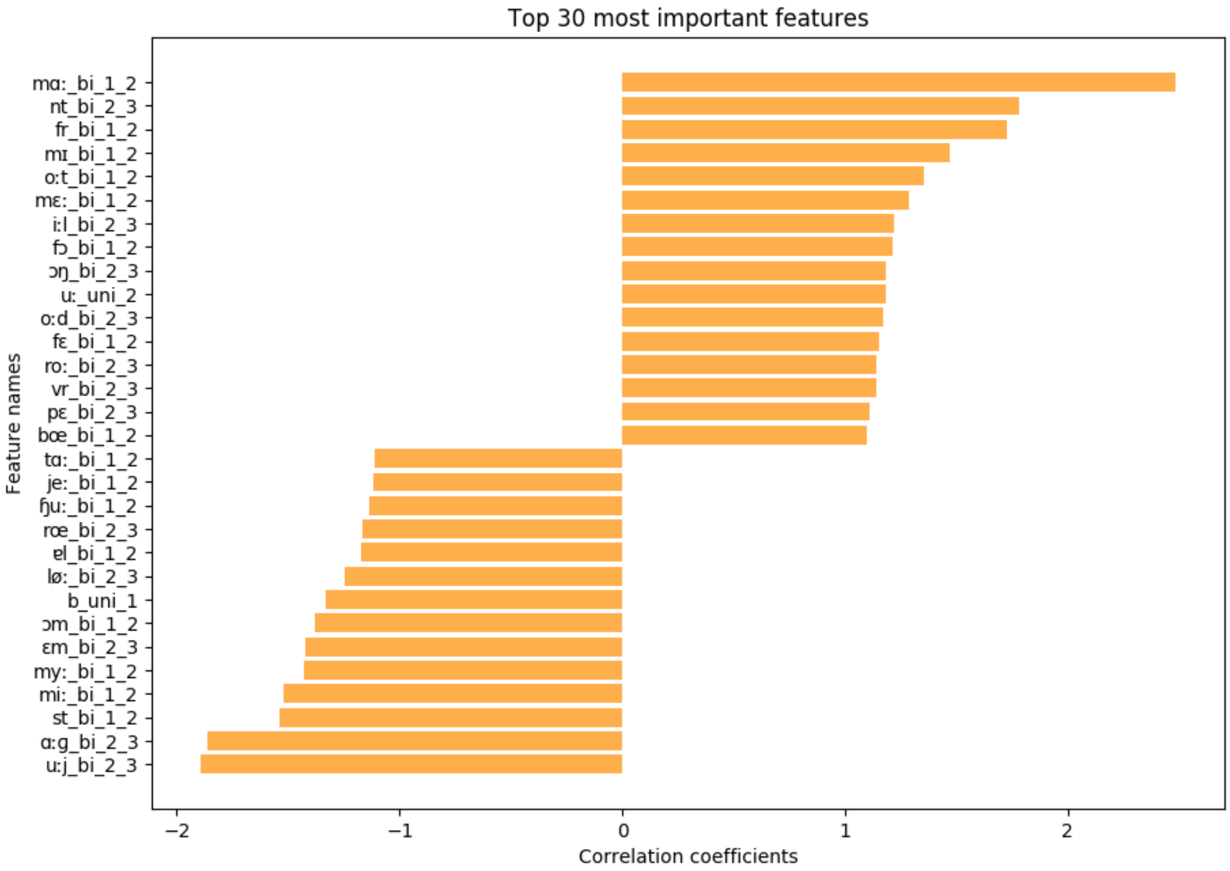

Figure 11: Top 30 features to classify HC and SCI group selected by Sparv-ROC. The x-axis shows the correlation coefficients of the feature and the y-axis shows the name of the feature. The longer the orange bar is, the higher contribution the feature makes. All bars to the left indicate negative classes and to the right positive classes.

Figure 11 shows the top 30 features that contribute most to classify HC and SCI group selected by Sparv-Acc strategy. The negative values indicate the negative class which in this case is HC, and the positive values contribute to the SCI class.

The top features for SCI group were: ‘bœ_bi_1_2’, ‘pɛ_bi_2_3’, ‘vr_bi_2_3’, ‘roː_bi_2_3’, ‘fɛ_bi_1_2’, ‘o:d_bi_2_3’, ‘uː_uni_2’, ‘ɔŋ_bi_2_3’, ‘fɔ_bi_1_2’, ‘i:l_bi_2_3’, ‘mɛː_bi_1_2’, ‘o:t_bi_1_2’, ‘mɪ_bi_1_2’, ‘fr_bi_1_2’, ‘nt_bi_2_3’ and ‘mɑː_bi_1_2’.

The top features for HC were: 'u:j_bi_2_3', 'ɑ:g_bi_2_3', 'st_bi_1_2', 'miː_bi_1_2', 'myː_bi_1_2', 'ɛm_bi_2_3', 'ɔm_bi_1_2', 'b_uni_1', 'løː_bi_2_3', 'ɐl_bi_1_2', 'rœ_bi_2_3', 'ɧuː_bi_1_2', 'jeː_bi_1_2' and 'tɑː_bi_1_2'.

All top 30 features came from phonological groups where [mɑː] in the beginning of words occurred more frequent in SCI group and words having [u:j] at the second and third place was the strongest predictor for HC.

13 SCI predictors and 9 HC predictors were identical to the output of the Sparv-Acc group. Words beginning with [miː], [tɑː], [ɔm] and words having [ɑ:g] in the middle were top HC features in all models; words beginning with [fr], [mɑː], [mɛː], [mɪ] and [o:t], words with [uː] at the second position and words with [i:l], [nt], [pɛ] at the second and third position were common indicators for SCI subjects.

Feed forward neural networks

The following chart demonstrates the statistics of 12 models built upon features described in the last section for HC vs SCI:

Table 13: The validation statistics of NN models in respect to the feature elimination strategy for HC vs SCI individual classification

Model	Validation accuracy mean	SD	Matthews correlation coefficient mean	ROC AUC mean	SD	Feature selection measurement	Annotation tool
M1	70.03	7.72	0.37	0.69	0.08	ROC AUC c	Parsey
M2	73.00	8.00	0.32	0.66	0.1	ROC AUC	Parsey
M3	74.34	6.48	0.37	0.69	0.11	ROC AUC	Parsey
M4	69.45	9.42	0.34	0.68	0.1	ROC AUC	Sparv
M5	73.68	6.19	0.34	0.67	0.09	ROC AUC	Sparv
M6	72.34	8.10	0.32	0.65	0.09	ROC AUC	Sparv
M7	**71.62**	**7.65**	**0.40**	**0.71**	**0.07**	**Accuracy**	**Parsey**
M8	75.30	7.87	0.40	0.70	0.11	Accuracy	Parsey
M9	75.44	7.64	0.40	0.70	0.11	Accuracy	Parsey
M10	69.39	9.51	0.38	0.70	0.09	Accuracy	Sparv
M11	71.12	9.26	0.36	0.68	0.11	Accuracy	Sparv
M12	71.28	10.21	0.34	0.67	0.08	Accuracy	Sparv

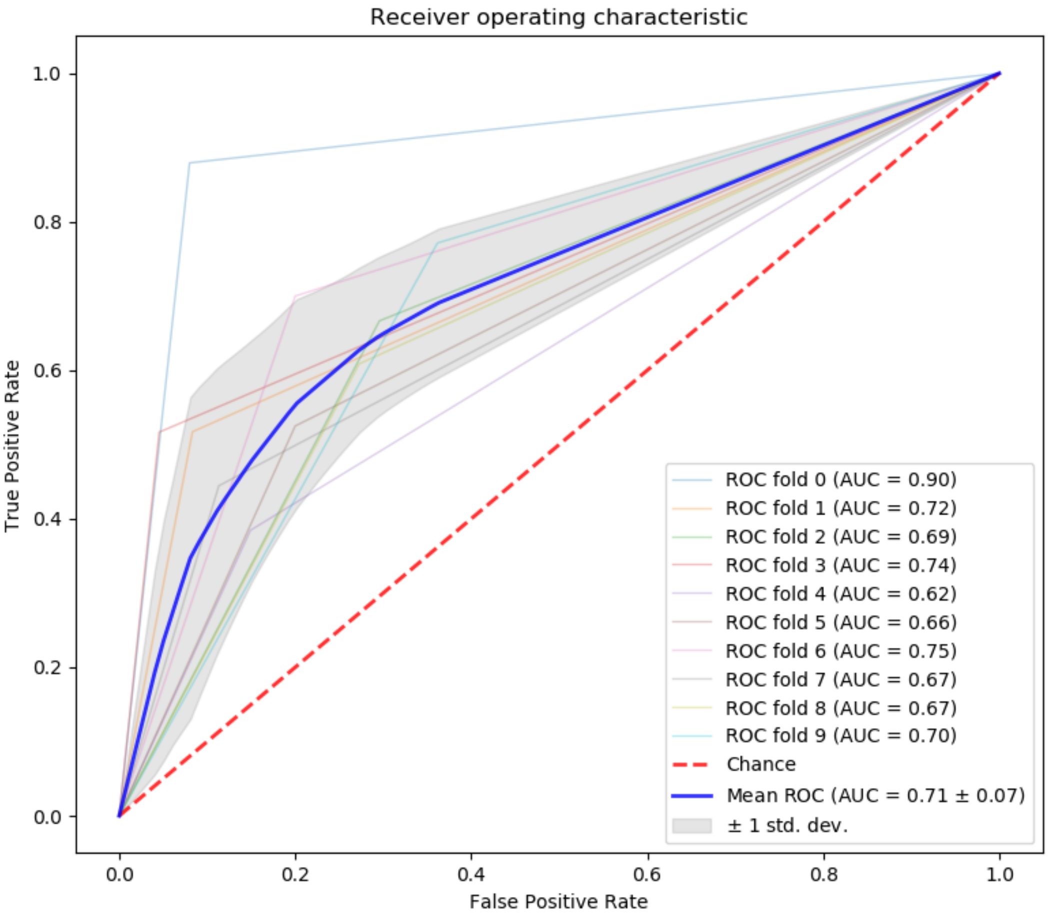

Figure 12: Mean ROC curve and AUC of the 10-fold cross validation model M7 for HC vs SCI classification. The x-axis shows the false positive rate and the y-axis shows the true positive rate of the evaluation. The 10 light color curves correspond to the ROCs for each one of the 10 folds. The bold blue curve shows the mean ROC and the grey shadow shows the standard deviation of the mean ROC curve. The red dotted line represents the baseline that predicts the class by chance. Curves equal or below the red dotted line indicate a bad model.

The most comprehensive NN model was M7 which employed features selected by Parsey-Acc method. Although the mean accuracy was not the highest (71.62%), it got the highest MCC (0.40) and ROC AUC mean with the lowest SD (0.71 +/- 0.07) among all models. When it comes to the area under the ROC value for each fold, the score was up to 0.90 and down to 0.62 which was higher than the baseline (0.5). Comparing the performance of feature selection, the accuracy was slightly raised by almost 4% and the MCC was increased by 28%.

Table 14 demonstrates the configuration of the M7 model. M7 had a very simple NN structure compared to the one for HC vs MCI individual detection. It did not have any hidden layer. The input layer reduced the data dimension from 150 (generated by Parsey-Acc approach to identify SCI individual from HC during feature selection) to 10 and used tanh for activation. The output layer had 3-dimension-input with Softmax as activation function.

Table 14: The configuration of the M7 model for HC vs SCI individual classification

Layer	Dimension	Activation	Dropout
Input layer	10 (input 150 dimensions)	tanh	N/A
Output layer	3	Softmax	N/A

Study 3: MCI vs SCI

Feature selection

Table 15-18 demonstrates the number of features and folds that were employed in each feature group corresponds to their validation MCC, validation ROC AUC and validation accuracy produced by the SVC classifier for MCI vs SCI individual classification. The composition of the table is the same as the ones in Study 1.

Text annotated using Parsey Universal, model selected by accuracy

Table 15: The number of features, folds and classification performance of feature groups using Parsey-Acc feature elimination strategy for MCI vs SCI individual classification.

Feature group	Number of features	Number of folds	MCC	ROC_AUC	Accuracy
Morphological features	2	6	0.07	0.487	55%
Syntactic features	2	5	0.01	0.495	51.8%
Phonological features	193	6	0.290	0.356	62.9%
Other features (Dialog event + syllabus features)	5	6	0.303	0.353	66.8%
Morphological + Phonological + Other features	98	5	0.378	0.311	69.7%

The model that outperformed was trained with group 5-fold cross validation by Parsey-Acc approach. Morphological, phonological and dialog features consisted of the best model from feature extraction including count of self-correction, MLU, adjectives ratio, participle ratio, 10 unigram positional segment frequencies and 84 bigram positional segment frequencies.

We reached 55% accuracy, 52% accuracy, 63% accuracy and 67% accuracy with morphological features, syntactic features, phonological features and other features, respectively. The accuracy was increased to 69.7% combining less than half of the phonological features, all morphological features, dialog features and MLU with 0.378 MCC and 0.311 ROC AUC.

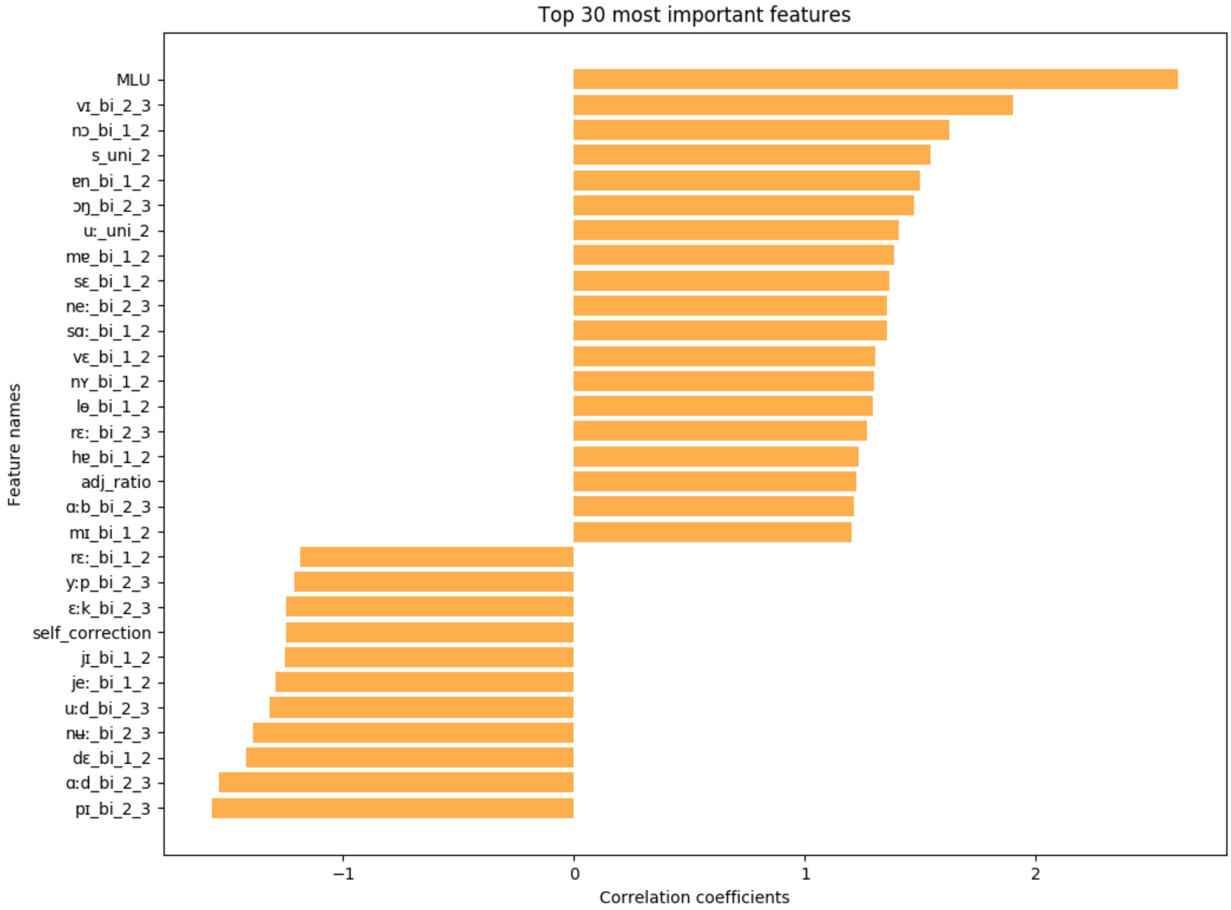

Figure 13: Top 30 features to classify MCI and SCI group selected by Parsey-Acc. The x-axis shows the correlation coefficients of the feature and the y-axis shows the name of the feature. The longer the orange bar is, the higher contribution the feature makes. All bars to the left indicate negative classes and to the right positive classes.

Figure 13 shows the top 30 features that contribute most to classify MCI and SCI group selected by Parsey-Acc strategy. The negative values indicate the negative class which in this case is MCI, and the positive values contribute to the SCI class.

The top MCI features were: 'pɪ_bi_2_3', 'ɑːd_bi_2_3', 'dɛ_bi_1_2', 'nʉː_bi_2_3', 'uːd_bi_2_3', 'jeː_bi_1_2', 'jɪ_bi_1_2', count of self-correction, 'ɛːk_bi_2_3', 'yːp_bi_2_3' and 'rɛː_bi_1_2'. Words with [pɪ] in the middle were the strongest predictor for MCI.

The top SCI features were: 'mɪ_bi_1_2', 'ɑːb_bi_2_3', adjective ratio, 'hɐ_bi_1_2', 'rɛː_bi_2_3', 'lɵ_bi_1_2', 'nʏ_bi_1_2', 'vɛ_bi_1_2', 'sɑː_bi_1_2', 'neː_bi_2_3', 'sɛ_bi_1_2', 'mɐ_bi_1_2', 'uː_uni_2', 'ɔŋ_bi_2_3', 'ɐn_bi_1_2', 's_uni_2', 'nɔ_bi_1_2', 'vɪ_bi_2_3' and MLU. MLU was again the most important feature to identify SCI considering the models trained for HC vs. SCI task.

The proportion of top 30 features for MCI vs SCI classification was 0.58. Combining with the accuracies of feature groups, all selected morphological predictors were highly associated with SCI group. Self-correction was much more observed in MCI group and the mean MLU of SCI individuals was the highest among all groups. 'pɪ_bi_2_3' became the most important predictor since was only found in MCI samples.

Using the same annotated text with accuracy as benchmark, ‘nʉː_bi_2_3’, ‘pɪ_bi_2_3’ and ‘uːd_bi_2_3’ were the common predictors compared to the top 30 features for MCI individual in the HC vs. MCI model; MLU and ‘mɪ_bi_1_2’ were the common ones for SCI individual compared to the top 30 features in the HC vs. SCI model.

Text annotated using Parsey Universal, model selected by area under ROC and Matthews correlation coefficient

Table 16: The number of features, folds and classification performance of feature groups using Parsey-ROC feature elimination strategy for MCI vs SCI individual classification

Feature group	Number of features	Number of folds	MCC	ROC_AUC	Accuracy
Morphological features	3	5	0.082	0.530	53.2%
Syntactic features	2	6	0.046	0.514	50.1%
Phonological features	193	6	0.290	0.644	62.9%
Other features (Dialog event + syllabus features)	5	6	0.303	0.647	66.8%
Morphological + Phonological + Other features	113	6	0.385	0.699	69.6%

The 113 selected features for the model made by Parsey-ROC method was trained with group 6-fold cross validation, which contained count of interruptions, count of unclear segments, count of self-corrections, MLU, ASL, ratio of participle, 8 unigram PSFs and 99 bigram PSFs. The count of self-corrections, unclear segments and interruptions contribute to MCI individual detection, and MLU, participle ratio and ASL were indicators for SCI individuals.

Compared to the models trained with Parsey-Acc approach, we reached 66.8% accuracy, 0.303 MCC with dialog feature, MLU and ASL, which was the same but with much higher ROC AUC (increased from 0.353 to 0.647). Although the accuracies of models built with only morphological features, syntactic features and phonological features were lower than or equal to the ones from Parsey-Acc approach, their ROC AUC were significantly improved. By removing two morphological features and adding more phonological features, the combined-feature model gained overall better performance (69.6% accuracy, 0.385 MCC and 0.699 ROC AUC).

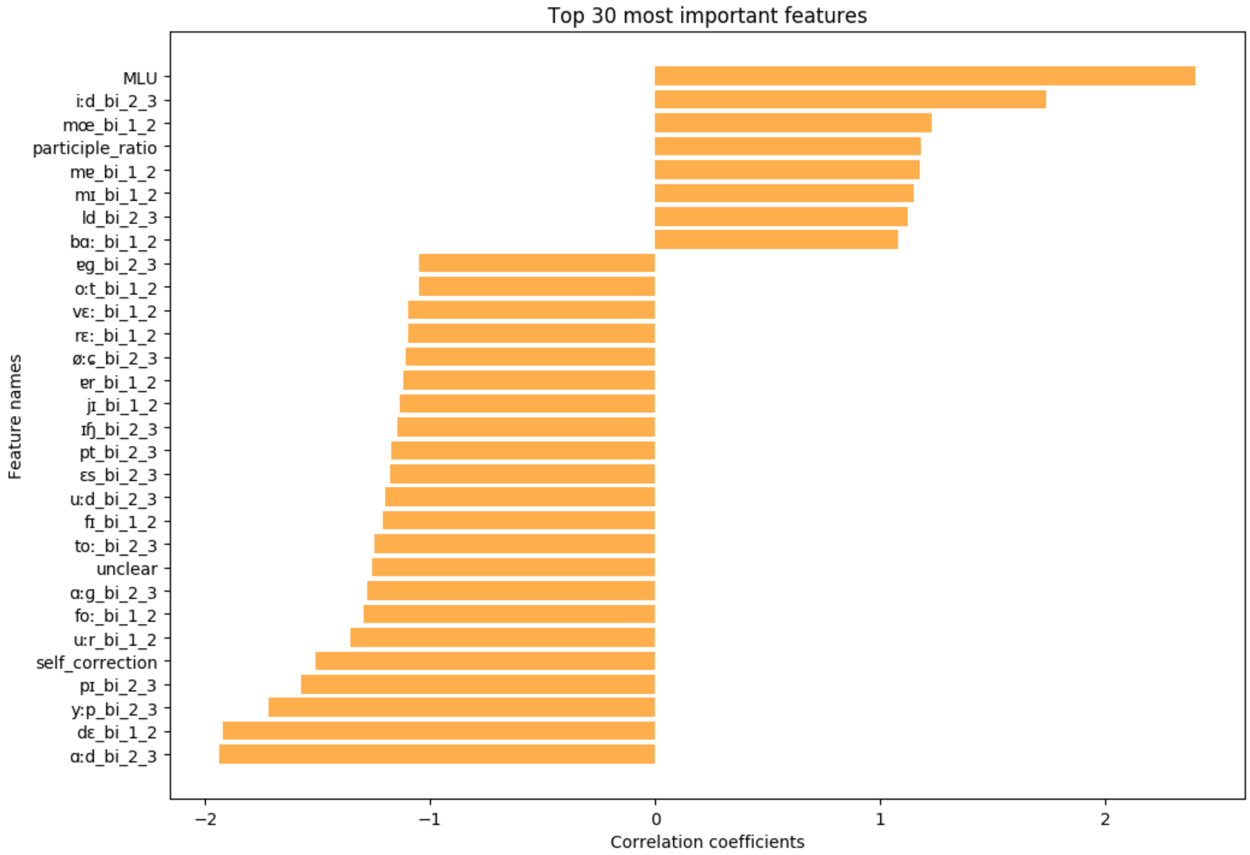

Figure 14: Top 30 features to classify MCI and SCI group selected by Parsey-ROC. The x-axis shows the correlation coefficients of the feature and the y-axis shows the name of the feature. The longer the orange bar is, the higher contribution the feature makes. All bars to the left indicate negative classes and to the right positive classes.

Figure 14 shows the top 30 features that contribute most to classify MCI and SCI group selected by Parsey-ROC strategy. The negative values indicate the negative class which in this case is MCI, and the positive values contribute to the SCI class.

The top features for MCI group were: ‘ɑ:d_bi_2_3’, ‘dɛ_bi_1_2’, ‘y:p_bi_2_3’, ‘pɪ_bi_2_3’, count of self-correction, ‘u:r_bi_1_2’, ‘fo:_bi_1_2’, ‘ɑ:g_bi_2_3’, ‘unclear’, ‘to:_bi_2_3’, ‘fɪ_bi_1_2’, ‘u:d_bi_2_3’, ‘ɛs_bi_2_3’, ‘pt_bi_2_3’, ‘ɪɧ_bi_2_3’, ‘jɪ_bi_1_2’, ‘ɐr_bi_1_2’, ‘ø:ɕ_bi_2_3’, ‘rɛ:_bi_1_2’, ‘vɛ:_bi_1_2’, ‘o:t_bi_1_2’ and ‘ɐg_bi_2_3’.

The top features for SCI group were: ‘bɑ:_bi_1_2’, ‘ld_bi_2_3’, ‘mɪ_bi_1_2’, ‘mɐ_bi_1_2’, ratio of participle, ‘mœ_bi_1_2’, ‘i:d_bi_2_3’ and MLU.

Among the top 30 features, 22 ones were MCI individual predictors including count of self-corrections, unclear segments and 20 bigram PSFs, where MLU, participle ratio and 6 bigram PSFs contributed to SCI group. It tells that MCI individuals tend to talk unclear and correct themselves more compare to individual with SCI.

Compared to the top 30 feature group of MCI vs. SCI model from text annotated by Parsey Universal but selected by accuracy, the intersection of two feature groups for MCI group were count of self-correction, words begin with [dɛ], [rɛ:] and [jɪ], words with [pɪ], [u:d], [ɑ:d] and [y:p] in the middle; the common predictors for SCI group were MLU, words having [mɐ] and [mɪ] in the front. The proportion of top 30 features for the 2 classes were dramatically different which implies that the feature selection benchmarks influenced the feature importance.

Converging the MCI predictors of HC vs MCI model trained by Parsey-ROC approach and the current model, bigram PSFs such as ‘oːt_bi_1_2’, ‘uːr_bi_1_2’ and ‘ɐɡ_bi_2_3’ were the common ones. Compared to the SCI predictors of HC vs SCI model trained by Parsey-ROC approach, words begin with [mɪ] was the common one which was also the universal SCI feature in models created by Parsey-Acc approach.

Text annotated using Sparv, model selected by accuracy

Table 17: The number of features, folds and classification performance of feature groups using Sparv-Acc feature elimination strategy for MCI vs SCI individual classification

Feature group	Number of features	Number of folds	MCC	ROC_AUC	Accuracy
Morphological features	9	5	0.094	0.536	53.4%
Syntactic features	2	5	0.01	0.505	51.8%
Phonological features	194	6	0.288	0.644	62.7%
Other features (Dialog event + syllabus features)	5	6	0.303	0.647	66.8%
Morphological + Phonological + Other features	124	5	0.368	0.682	69.4%

The most important feature group with the highest accuracy using Sparv-Acc method was selected by group 5-fold cross validation. The 124 features contained 1 dialog event, 1 syllabus feature, 1 syntactic feature, 7 POS features and 114 phonemic features (8 unigram PSFs and 106 bigram PSFs) where MLU, ASL, ratio of adverbs, particles, adjectives, participles and dash were indicators for SCI individuals and count of self-correction, ratio of prepositions and ratio of light verbs were indicators for MCI individuals. Most of the POS features were included in the top 30 feature set in the current classification task. However, the increased number of POS features did not make significant improvement compared to the models trained with Parsey Universal annotated text.

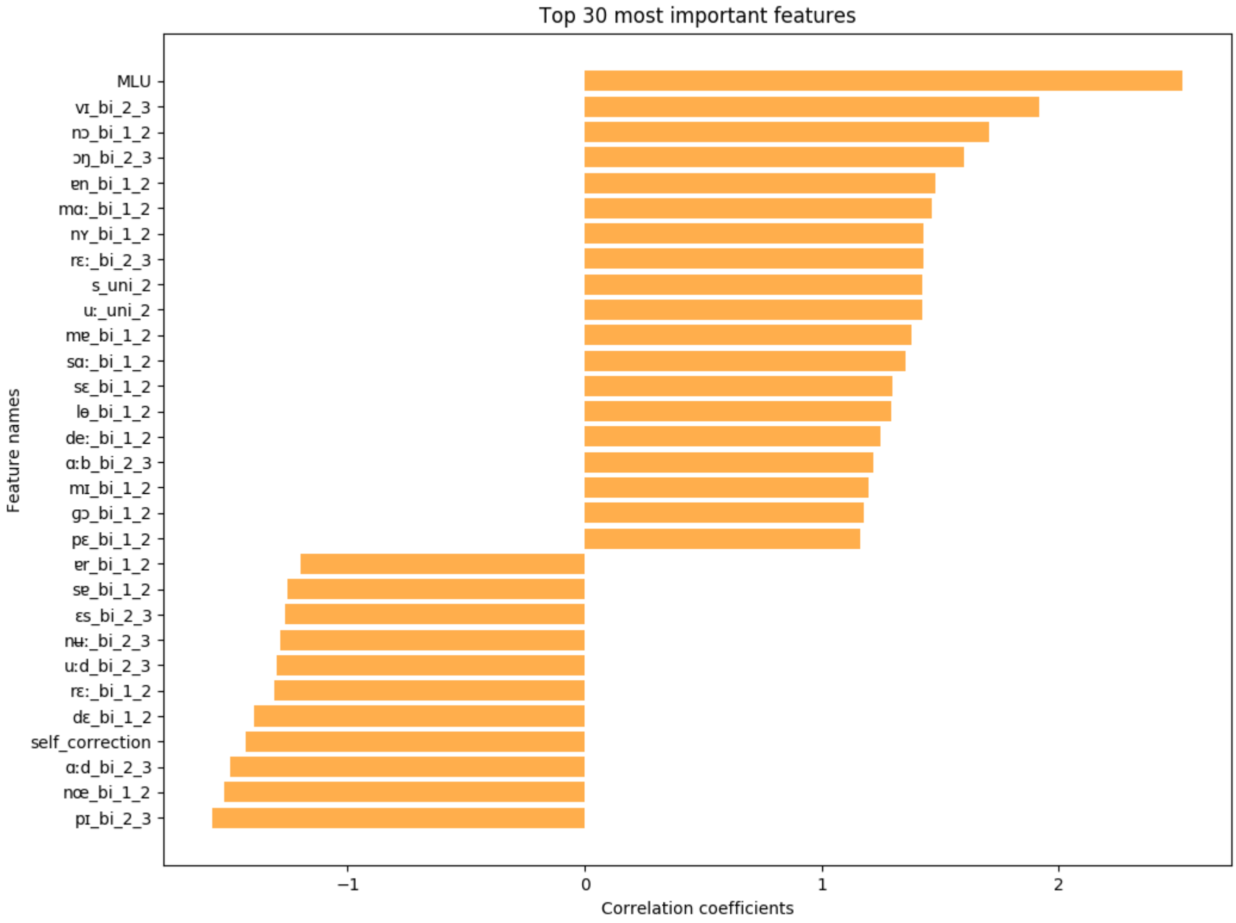

Figure 15: Top 30 features to classify MCI and SCI group selected by Sparv-Acc. The x-axis shows the correlation coefficients of the feature and the y-axis shows the name of the feature. The longer the orange bar is, the higher contribution the feature makes. All bars to the left indicate negative classes and to the right positive classes.

Figure 15 shows the top 30 features that contribute most to classify MCI and SCI group selected by Sparv-Acc strategy. The negative values indicate the negative class which in this case is MCI, and the positive values contribute to the SCI class.

The top features for MCI group were: ‘pɪ_bi_2_3’, ‘nœ_bi_1_2’, ‘ɑːd_bi_2_3’, count of self-correction, ‘dɛ_bi_1_2’, ‘rɛː_bi_1_2’, ‘uːd_bi_2_3’, ‘nʉː_bi_2_3’, ‘ɛs_bi_2_3’, ‘sɐ_bi_1_2’ and ‘ɐr_bi_1_2’.

The top features for SCI group were: ‘pɛ_bi_1_2’, ‘gɔ_bi_1_2’, ‘mɪ_bi_1_2’, ‘ɑːb_bi_2_3’, ‘deː_bi_1_2’, ‘lɵ_bi_1_2’, ‘sɛ_bi_1_2’, ‘sɑː_bi_1_2’, ‘mɐ_bi_1_2’, ‘uː_uni_2’, ‘s_uni_2’, ‘rɛː_bi_2_3’, ‘nʏ_bi_1_2’, ‘mɑː_bi_1_2’, ‘ɐn_bi_1_2’, ‘ɔŋ_bi_2_3’, ‘nɔ_bi_1_2’, ‘vɪ_bi_2_3’ and MLU.

Compared to the top 30 features of Parsey-Acc MCI vs. SCI model, 7 predictors were common in MCI individuals including count of self-correction, words begin with [dɛ], [rɛː] and words with [ɑːd], [uːd], [nʉː] and [pɪ] in the middle; MLU, words begin with [lɵ], [mɪ], [mɐ], [nɔ], [nʏ], [sɑː], [sɛ] and [ɐn], words with [uː] and [s] at the second position and words with [ɔŋ], [ɑːb], [rɛː] and [vɪ] in the middle consisted of the common predictors for SCI individuals.

Converging the MCI predictors of top 30 features in Sparv-Acc HC vs. MCI model and the ones in the current model, no common marker was found. However, when having similar comparison with SCI predictors in Spar-Acc HC vs. SCI model, 2 features consisted of the intersection which includes MLU, and 'mɪ_bi_1_2'. MLU was also one of the strongest SCI features in all Parsey-Acc generated related models (HC vs. SCI and MCI vs. SCI).

Text annotated using Sparv, model selected by area under ROC and Matthews correlation coefficient

Table 18: The number of features, folds and classification performance of feature groups using Sparv-ROC feature elimination strategy for MCI vs SCI individual classification

Feature group	Number of features	Number of folds	MCC	ROC_AUC	Accuracy
Morphological features	9	5	0.094	0.550	53.7%
Syntactic features	2	10	0.017	0.513	45.8%
Phonological features	245	10	0.243	0.646	61.2%
Other features (Dialog event + syllabus features)	5	6	0.303	0.647	66.8%
Morphological + Phonological + other features	150	5	0.414	0.705	71.6%

The combined-feature model using Sparv-ROC approach to identify MCI and SCI individuals was trained with 150 features and group 5-fold cross validation. Count of self-corrections, ASL, MLU, ratio of adjectives, dash, light verbs, particles, participles, prepositions and 141 phonological features (8 unigram PSFs and 133 bigram PSFs) consisted of the features that outperformed.

All models built in the present study performed better than the ones trained with the other feature selection strategies. With the greatest number of morphological features, phonological features and other features, the final model gained 71.6% accuracy, 0.414 MCC and 0.705 ROC_AUC. It turns out that Sparv-ROC is the best feature selection strategy to classify MCI subjects and SCI subjects.

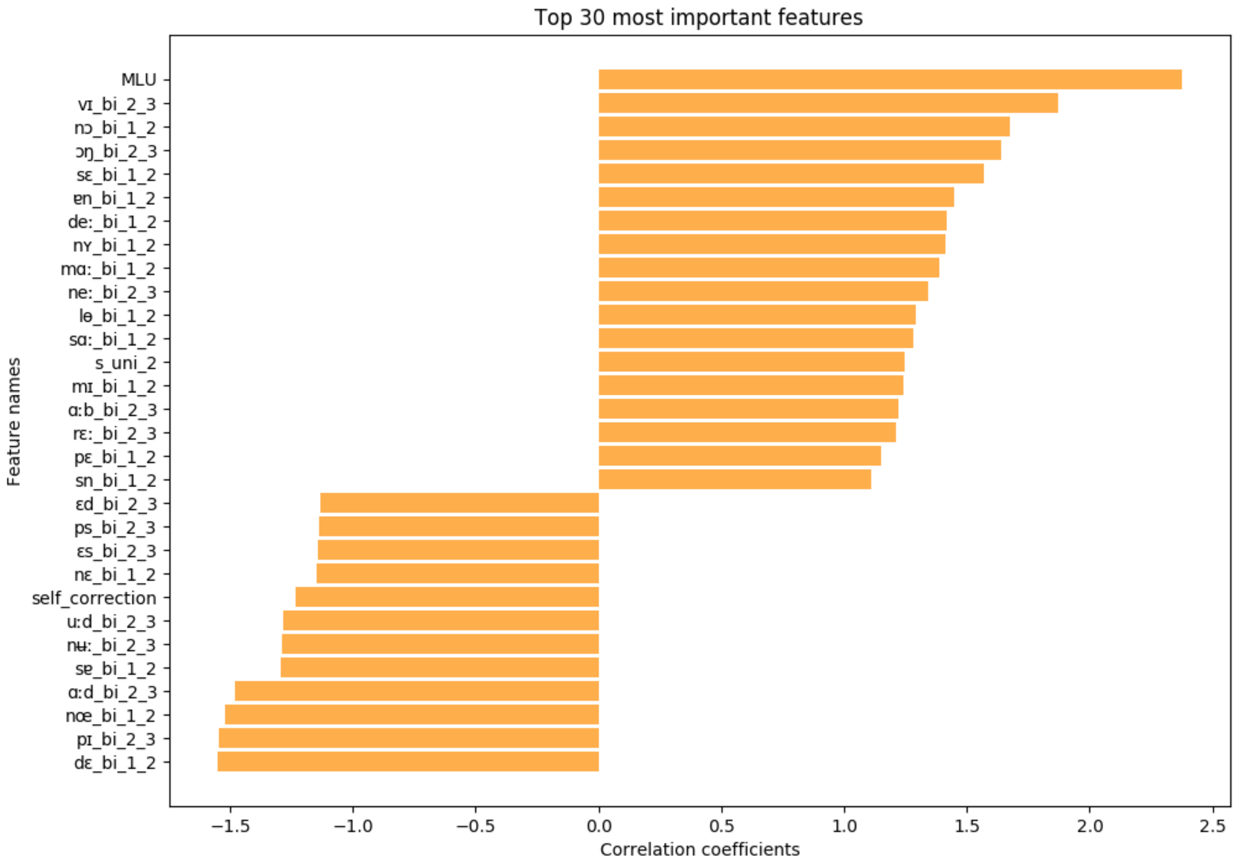

Figure 16: Top 30 features to classify MCI and SCI group selected by Sparv-ROC. The x-axis shows the correlation coefficients of the feature and the y-axis shows the name of the feature. The longer the orange bar is, the higher contribution the feature makes. All bars to the left indicate negative classes and to the right positive classes.

Figure 16 shows the top 30 features that contribute most to classify MCI and SCI group selected by Sparv-ROC strategy. The negative values indicate the negative class which in this case is MCI, and the positive values contribute to the SCI class.

The top features for MCI group were: 'dɛ_bi_1_2', 'pɪ_bi_2_3', 'nœ_bi_1_2', 'ɑːd_bi_2_3', 'sɐ_bi_1_2', 'nʉː_bi_2_3', 'uːd_bi_2_3', count of self-correction, 'nɛ_bi_1_2', 'ɛs_bi_2_3', 'ps_bi_2_3' and 'ɛd_bi_2_3'.

The top features for SCI group were: 'sn_bi_1_2', 'pɛ_bi_1_2', 'rɛː_bi_2_3', 'ɑːb_bi_2_3', 'mɪ_bi_1_2', 's_uni_2', 'sɑː_bi_1_2', 'lɵ_bi_1_2', 'neː_bi_2_3', 'mɑː_bi_1_2', 'nʏ_bi_1_2', 'deː_bi_1_2', 'ɐn_bi_1_2', 'sɛ_bi_1_2', 'ɔŋ_bi_2_3', 'nɔ_bi_1_2', 'vɪ_bi_2_3' and MLU.

The most important predictor for MCI individual is 'dɛ_bi_1_2'. Ratio of prepositions and light verbs also contributed to identify MCI individuals with lower feature importance. Count of self-corrections was a strong MCI feature in all MCI vs. SCI models' top 30 feature set. The most important predictor for SCI individuals in this case was MLU which was included in all models that identify SCI individuals.

Comparing the top 30 features of Sparv-ROC MCI vs. SCI model with Parsey-ROC MCI vs. SCI model, MLU and words begin with [mɪ] were the common features to identify SCI subjects; Count of self-correction, words begin with [dɛ], words with [ɑːd], [uːd], [pɪ] in the middle were the intersection for MCI individuals. Converging Sparv-ROC MCI vs. SCI model and Sparv-Acc MCI vs. SCI model, MLU, words begin with [deː], [lɵ], [mɑː], [mɪ], [nɔ], [nʏ], [pɛ], [sɑː], [ɐn], words with [neː], [ɑːb], [vɪ], [ɔŋ] in the middle, and words with [s] at the second position were the common ones for SCI group in the top 30 feature sets; Count of self-corrections, words begin with [dɛ], [nœ], [sɐ] and words with [ɑːd], [uːd], [ɛs], [nʉː], [pɪ] in the middle were intersection for MCI group.

When it comes to detecting SCI subject with Sparv-ROC approach, 'mɑː_bi_1_2', 'mɪ_bi_1_2', 'ɔŋ_bi_2_3' were the common predictors in HC vs. MCI and MCI vs. SCI models. For MCI subjects, 'nʉː_bi_2_3' is the only intersection converging the top 30 features of HC vs. SCI and MCI vs. CI models.

In addition to the common top features, we found that words begin with [uːr] and words with [pɪ] in the middle only exist in MCI samples and words begin with [nʏ] only exist in SCI samples.

Feed forward neural networks

The following chart demonstrates the statistics of 12 models built upon features described in the last section to distinguish MCI and SCI individuals:

Table 19: The validation statistics of NN models in respect to the feature elimination strategy for MCI vs SCI individual classification

Model	Validation accuracy mean	SD	Matthews correlation coefficient mean	ROC AUC mean	SD	Feature selection measurement	Annotation tool
M1	71.78	3.76	0.39 (K3)	0.69	0.06	ROC AUC	Parsey
M2	72.77	3.45	0.39 (K3)	0.67	0.05	ROC AUC	Parsey
M3	70.80	4.39	0.38 (K3)	0.68	0.05	ROC AUC	Parsey
M4	71.29	3.60	0.37 (K3)	0.68	0.04	ROC AUC	Sparv
M5	70.55	3.17	0.37 (K3)	0.68	0.05	ROC AUC	Sparv
M6	71.18	4.96	0.39 (K3)	0.69	0.04	ROC AUC	Sparv
M7	74.91	9.8	0.38 (K10)	0.72	0.1	Accuracy	Parsey
M8	**75.33**	**8.92**	**0.39 (K10)**	**0.71**	**0.07**	**Accuracy**	**Parsey**
M9	73.75	14.20	0.38 (K10)	0.71	0.1	Accuracy	Parsey
M10	70.89	8.26	0.4 (K5)	0.70	0.07	Accuracy	Sparv
M11	74.82	10.45	0.36 (K10)	0.70	0.1	Accuracy	Sparv
M12	72.03	2.70	0.38 (K3)	0.68	0.05	Accuracy	Sparv

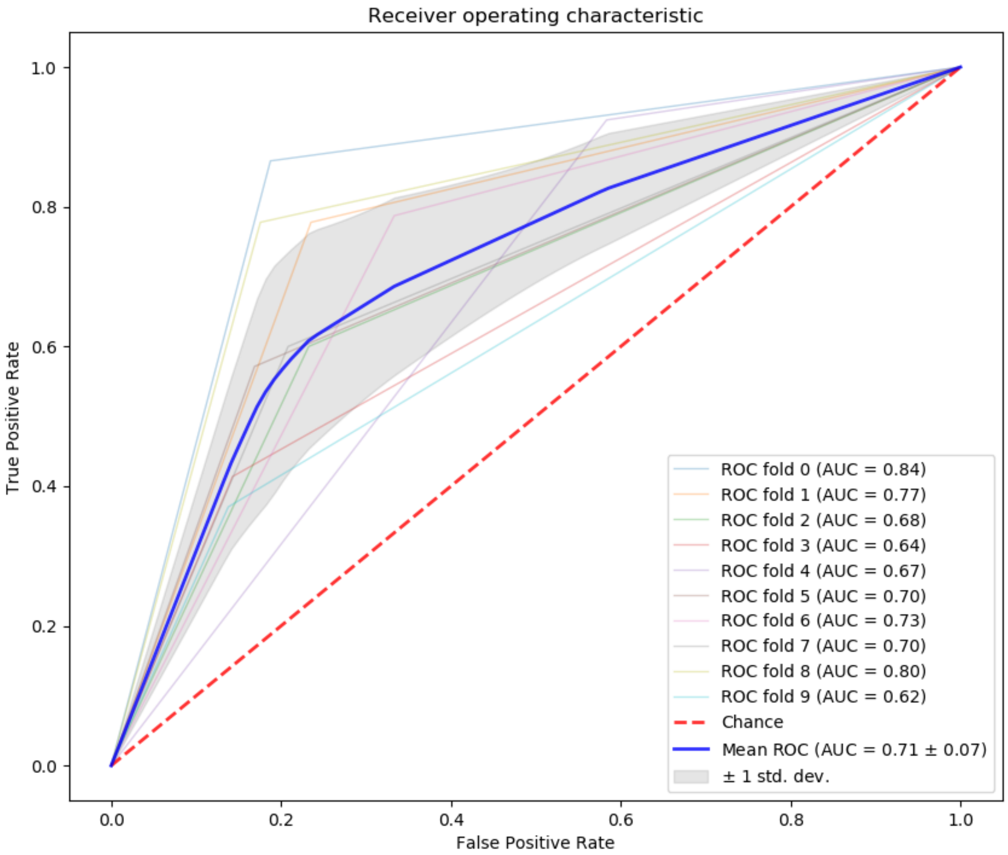

Figure 17: Mean ROC curve and AUC of the 10-fold cross validation model M8 for MCI vs SCI classification. The x-axis shows the false positive rate and the y-axis shows the true positive rate of the evaluation. The 10 light color curves correspond to the ROCs for each one of the 10 folds. The bold blue curve shows the mean ROC and the grey shadow shows the standard deviation of the mean ROC curve. The red dotted line represents the baseline that predicts the class by chance. Curves equal or below the red dotted line indicate a bad model.

The best NN model to classify MCI and SCI individuals was M8 which also came from text annotated using Parsey Universal and features selected by accuracy. It had the highest mean accuracy and mean ROC AUC among the 12 models. Although its mean MCC was only improved by 3%, the accuracy was increased from 69.7% to 75.3% and the value of ROC AUC got raised by 56%. Throughout 10-fold cross validation, the highest ROC AUC score reached 0.84 at fold 0, and the lowest one was 0.62 at fold 9.

Table 20 shows the configuration of the M8 model which only has one hidden layer. It took 98 dimensions of vectors as input (the 98 top features generated by Parsey-Acc approach to differentiate MCI and SCI individuals during feature selection), passed through tanh function and dropped out 40% of activation output. The hidden layer connected 182 columns with 40 neurons. It applied Relu for activation and 0.3 dropout. In the end, the output layer reduces the output of the hidden layer to three and uses Softmax activation for the final computation.

Table 20: The configuration of the M8 model for MCI vs SCI individual classification

Layer	Dimension	Activation	Dropout
Input layer	182 (input 98 dimensions)	tanh	0.4
Hidden layer	40	ReLu	0.3
Output layer	3	Softmax	N/A

Discussion

The present study accomplished three classification tasks on distinguishing different stages of cognitive decline using natural language processing and machine learning techniques. During the study, we analyzed speech transcriptions of Swedish participants who were diagnosed with mild cognitive impairment (MCI) or subjective cognitive impairment (SCI) in clinic and investigated linguistic predictors that were not explored in previous studies. Although spontaneous speech provides clinical data that are tailored to individual productions, extracting effective information for general analysis can be a challenging task. By understanding the structure of the data and the background of data collection, we are providing valuable information using different annotation tools and automatic phonetic transcription. In this context, Parsey Universal Server was employed to annotate connected speech in Swedish individuals with SCI, MCI, and healthy controls. The proposed method can be adapted to other languages as well and will be available to the public as a portable solution that can be deployed in all major operating systems. The two annotation tools were employed and compared and, 689 dimensions of features were extracted from the annotated data. A novel feature engineering method has been proposed to identify the best composition of features for the three classification tasks. Further, the beneficial result of the method and its selected predictors were maximized by feed forward neural network models.

There were very few studies applied both NLP and machine learning techniques on SCI detection since the score of neuropsychological tests of people with SCI are close to normal and the severity of SCI are often overlooked. Through our findings, phonological predictors such as positional biphone frequency was the key predictor to detect SCI individuals and MCI individuals. Dialog related features, morphological features and mean length of utterance could be employed to differentiate MCI and SCI individuals (Themistocleous, Eckerström, et al., 2018; Themistocleous et al., 2020), but the most important features were self-corrections, mean length of utterance, and biphone positional segment frequencies. Specifically, individuals with MCI produced more long vowels than SCI ones, more unrounded vowels than rounded ones, more stops follow by back vowels, and did more self-corrections in spontaneous speech. SCI individuals tended to produce longer utterances and more nasal consonants follow by close front vowels than MCI ones.

Syntactic complexity of MCI and SCI individuals

Mean length of utterance (MLU) was reported as strong predictor identifying healthy controls from individuals with MCI by Beltrami et al. (2016) in topic description tasks. MLU is shorter in patients with Alzheimer's disease (CAN & Gülmira, 2018) and dementia (Fraser et al., 2014). While MLU was employed in the final neural network model to detect people with MCI from healthy controls, it was a strong predictor for SCI individuals rather than healthy controls since SCI produced the longest utterances on average in the present study. In other studies, it was syntactic complexity was generally contrarily associated with the severity of cognitive impairment in previous studies (Hernández-Domínguez et al., 2018). As MLU is a marker of syntactic complexity, individuals with SCI were expected to talk shorter than healthy elderlies. However, the observation corresponds to the finding by Lundholm Fors et al. (2018) who applied syntactic analysis on the same transcriptions, even though MLU was not considered as a significant feature. Another significant syntactic marker to identify healthy people from MCI ones was mean length of word (MLW), which indicated that MCI individuals tend to produce simpler words. Mean word length was employed to discriminate healthy people from the ones with semantic dementia and progressive non-fluent aphasia in the work published by Fraser et al. (2014).

The present findings demonstrate that MLU and MLW can identify patients with MCI from healthy people. Nevertheless, other syntactic features, such as average dependency distance were excluded during elimination due to low correlation coefficient values, which is opposite to the related study by Beltrami et al. (2016), but similar as the output of Lundholm Fors et al. (2018). Since not all suggested syntactic complexity markers in Kokkinakis et al. (2017) were investigated, more features could be explored in future works.

Phonemic production of MCI and SCI individuals

Individuals with MCI or SCI do have noticeable phonological patterns during picture description tasks. More than 96% of the features that contributed to the final models for all classification tasks were phonological ones, which is consistent to findings by several studies (Gosztolya et al., 2019, 2016; Tóth et al., 2015).

To summarize the preferred phonemic patterns in MCI and SCI individuals, we extracted the intersections of features that contributed to all the best models during feature selection and look up the patterns with different categories of consonants (voiced plosives, voiceless plosives, nasal, fricatives, approximants, rhotic, labial, dental/alveolar, dental, palatal, velar, glottal, stop and sonorants) and vowels (front, central, back, close, close-mid, open-mid, open, long, short, rounded, unrounded) summarized in Swedish phonology (Engstrand, 2004). The evaluation method was inspired by Beland (1999), which investigated the phonological features of a French patient who was impaired in writing but had normal oral production by dividing phonemic categories into voiced/voiceless, nasal/non-nasal, continuant/non-continuant and rounded/unrounded. In terms of vowels, we analyzed the long-short vowel ratio, unrounded-rounded vowel ratio, front-back vowel ratio, frequency of open vowels, open-mid vowels, central vowels, close-mid vowels and close vowels. For consonants, the frequency of all categories was computed. The alpha value was set .05 to determine significant difference between the categories; that is, when the gap between two values for the same category is higher than 0.05, we consider that there is significant difference, else we considered the categories as being the same. Based on this criterion, we demonstrate and discuss the phonemic impairment and frequent patterns in MCI and SCI individuals in contrast to different references.

Phonemic impairment and frequent patterns in MCI individuals

Compared to healthy people, MCI individuals produced lower or less:

- long-short vowel ratio (MCI: 0.75, HC: 3),
- open vowels (frequency MCI: 0%, HC: 5%),

but higher or more:

- unrounded-rounded vowel ratio (MCI: 3.6, HC:1),
- open-mid vowels (frequency MCI: 11%, HC: 6%).

However, no consonant category was identified as significant.

Compared to SCI individuals, MCI ones tended to produce less:

- nasal (frequency: MCI: 4%, SCI: 11%)

but higher or more:

- long-short vowel ratio (MCI: 2, SCI: 0.69)
- unrounded-rounded vowel ratio (MCI: 5, SCI: 1.42),
- stop (frequency: MCI: 14%, SCI: 8%).

Compared to healthy people and SCI individuals, MCI ones were impaired in producing:

- open-mid vowels (frequency MCI: 6%, HC and SCI: 12%),
- open vowels (frequency MCI: 0%, HC and SCI: 5%),

- nasal (frequency MCI: 0%, HC and SCI: 18%),
- dental-alveolar (frequency MCI: 14%, HC and SCI: 23%),
- fricatives (frequency MCI: 2%, HC and SCI: 14%),
- labial (frequency MCI: 7%, HC and SCI: 14%),

but generated more or higher:

- long-short vowel ratio (MCI: 1.3, HC and SCI: 0.52),
- unrounded-rounded vowel ratio (MCI: 16, HC and SCI: 3),
- stop (frequency MCI: 18%, HC and SCI: 5%),

Phonemic impairment and frequent patterns in SCI individuals

Compared to healthy people, SCI individuals generated lower or less:

- long-short vowel ratio (SCI: 1.07, HC: 2),
- front-back vowel ratio (SCI: 1.27, HC: 2)

whereas produced more:

- central vowels (frequency: SCI: 5%, HC: 0%),
- open-mid vowels (frequency: SCI: 15%, HC: 8%),
- unrounded-rounded vowel ratio (SCI: 5.5, HC: 2).

Compared to MCI individuals and HC, SCI ones generally produced much less:

- close-mid vowels (frequency: SCI: 0%, MCI and HC: 13%),
- long-short vowel ratio (SCI: 0.5, MCI and HC: 1.89),
- voiceless plosives (frequency: SCI: 0%, MCI and HC: 17%),
- stop (frequency: SCI: 7%, MCI and HC: 17%),
- fricatives (frequency: SCI: 0%, MCI and HC: 7%),
- dentals (SCI: 7%, MCI and HC: 17%),

but more:

- open-mid vowels (frequency: SCI: 18%, MCI and HC: 13%),
- back vowels (frequency: SCI: 9%, MCI and HC: 4%),
- nasal (frequency: SCI: 13%, MCI and HC: 8%),
- dental-alveolar (frequency: SCI: 13%, MCI and HC: 8%),
- sonorants (frequency: SCI: 13%, MCI and HC: 8%),
- velar (frequency: SCI: 13%, MCI and HC: 4%),
- rhotic (frequency: SCI: 7%, MCI and HC: 0%),
- palatal (frequency: SCI: 7%, MCI and HC: 0%)

It turns out that individuals with SCI did not show significant difference from healthy people compared to individuals with MCI when it comes to phonology. This finding explains why phonemic features did not gain the highest performance during feature selection for the classification MCI vs SCI, compared to the other tasks. Moreover, we found that both individuals with MCI and SCI produced more short vowels than long vowels, back vowels than front vowels, unrounded vowels than rounded vowels, more open-mid vowels and less close vowels, compared to healthy people.

The comparison of long-short vowel ratio is HC > MCI subjects > SCI subjects. Recalling studies that analyzed phonemic markers on individuals with cognitive impairment, this finding is opposite to the work by Jarrold et al. (2014) and Themistocleous, Kokkinakis, et al. (2018), where longer vowels were suggested associate to longer speech time and longer preparation time to utter, and can be a sign of cognitive impairment(Charalambos & Bronte, 2018). The reason of the conflict findings might be that the current observation was based on vowels and consonants positioned at the first to the third index. However, long vowels were still more frequent in patients with MCI than patients with SCI, which implies that patients with MCI are more impaired (slower than SCI ones) in speech. Another phonemic finding was that patients with MCI gained the highest unrounded-rounded vowel ratio compared to SCI ones and healthy controls (MCI subjects > SCI subjects > HC). We argue that patients with MCI produce more unrounded vowels than rounded vowels in the beginning of words. In addition, the frequency of open-mid vowels was the highest in SCI individuals (SCI subjects > MCI subjects > HC) which correspond to the high occurrence of [ɛ] in the top features for SCI ones. Our findings show that MCI and SCI individuals employed words that begin with [m] such as [mɛː] and the words begin with [mɪ], [mɐ] were strong predictors for SCI individuals, but by analyzing [m] productions from text transcripts we cannot conclude that they constitute a significant marker that differentiates cognitive impairment from healthy people without checking with the audio transcripts.

The findings that evaluate phonological production in MCI using machine learning are novel and fill a gap in the existing literature to identify MCI and SCI subjects with positional segments frequency, such as, MCI individuals produced more stops but show decline in open vowels and labials while SCI individuals produced less close-mid vowels, fricatives, stops, voiceless plosives and dentals but more dental-alveolar. Related work by Lundeborg et al. (2015) adopted voice onset time value to determine phonological impairment in Swedish children. Lundeborg et al. (2015) argued that *when speaking rate is lower, the time interval between the release of the oral closure and the onset of voicing for both voiceless stops and the following vowel increase*. This pattern might also relate to longer speech time and longer preparation time for the next speech, which has similar impact as long vowels for MCI subjects (Basilakos, 2016).

Dialog character in MCI individuals and SCI individuals

During the study design, we marked most words containing symbol '-' as self-correction (except "t-shirt") which includes repetitions, rephrasing with word "eller", falsestarts and interruptions. These markers can be categorized as speech dysfluency markers and were employed in related studies to identify people with cognitive impairments (Croot et al., 2000; Fraser, Lundholm Fors, Eckerström, et al., 2019; Lundholm Fors et al., 2018; Pakhomov et al., 2010). We found that MCI individuals corrected themselves. Similar findings were provided by Lundholm Fors et al. (2018) who showed that increased occurrence of false starts in MCI individuals and increased number of interruptions in SCI individuals, followed by Fraser et al. (2019) who combined the finding with other language, speech, eye-movement and comprehension features to predict MCI individuals.

POS production in MCI individuals and SCI individuals

Individuals with MCI produced fewer infinitive verbs than healthy controls. In Swedish, infinitives usually appear after modal verbs (need, must, want, etc.), or verbs with a particle. This finding seems conflicting with Lundholm Fors et al. (2018) where the proportion of non-finite verb in main clause in MCI individuals was higher than healthy controls and SCI ones, and Hansson et al. (2000) who argued that increased proportion of infinitives can be a sign of language impairment. Individuals with SCI produced more adjectives and participles than MCI individuals and healthy controls. Several studies investigated the different utilization of adjectives of people with cognitive impairments against healthy elderlies (Alegria et al., 2013; Beltrami et al., 2016; Fraser et al., 2014; Gosztolya et al., 2019; Hernández-Domínguez et al., 2018; Jarrold et al., 2014). These findings on adjectives and participles can be associated with the syntactic complexity of in the present study that SCI people produced the longest utterances on average, whereas

MCI individuals uttered the shortest. Nevertheless, fewer adjectives and participles demarcated MCI individuals and was employed in the classification tasks to identify the ones with MCI.

Utility of the annotation tool, automatic PSF extractor, SVC-RRFE

An added contribution of present study was the development of tools and methods, such as jesdoit/parsey-server, the automatic positional segments frequency (PSF) extractor, and SVC-Randomized Recursive Feature elimination (SVC-RRFE) process that can be applied in multiple computational scenarios. The annotation tool (*Jesdoit/Parsey-Universal-Server*, 2018/2018) employed the Google's syntaxnet (Alberti et al., 2017) and provides computational linguistic analysis of Part-Of-Speech and syntactic relations of words in any languages in a portable and scalable manner. The automatic positional segments extractor contributes to computational methods of analyzing phonological production of individuals with MCI and SCI and can be expanded to n-gram PSF analysis. Combining binary search and SVC-RFE to reduce the steps of feature elimination provides another method to feature engineering. In clinical practice, the overall code of the present study automates the process of fetching morphological, syntactic, phonological and dialog character of individuals with MCI and SCI for physicians, neuropsychologists and speech-language pathologists. The top feature visualizer of the code assists clinicians to identify objective markers to diagnose patients with MCI and SCI, and the observed linguistic predictors can be employed to estimate the impact of applied treatments or detect the natural degenerative process of the condition. All functions in the code can expand to explore more markers for MCI and SCI individuals. They can also be utilized on other conditions that have language impairment background.

The performance of the two annotation tools, Sparv and Parsey Universal, was compared using the prediction models of two stage classifications during data processing. While Sparv outperformed in HC vs MCI and MCI vs SCI classifications, annotated data from Parsey Universal were enabled an improved prediction of individuals with SCI. Parsey Universal annotated data and accuracy selected features made up the better ones in all tasks, based on the observation. Also, data from Parsey Universal resulted in improved classification accuracy using feed-forward neural networks than data from Sparv. It turned out that the feature groups that were identified as the best among other selection methods do not always perform well in feed-forward neural network approach. Importantly, more features not always led to better performance. For example, all the chosen set of predictors were less than half of the original size - 689; Parsey-ROC method reached 69.1% accuracy for HC vs SCI classification using less predictors than Sparv-acc approach.

Comparing model predictions with clinical evaluation result through 10-fold cross validation feed-forward neural networks, we reached 76% mean accuracy, 73% mean ROC AUC to detect MCI subjects with 160 features including morphological, phonological, dialogue and syntactic features; For SCI subjects, we archived 71% mean accuracy, 71% mean ROC AUC with 150 features (phonological, dialogue, syntactic); For MCI vs SCI classification, we gained 75% mean accuracy and 71% mean ROC AUC with 98 features with morphological, phonological, dialogue and syntactic predictors. The highest validation accuracy for the three models were 83%, 79% and 84%, respectively, resulting in better performance than previous study (Lundholm Fors et al., 2018) given the same speech transcriptions.

To summarize, neural network models proved that they are an effective approach that identifies MCI/SCI individuals from HC and people with MCI from people with SCI given small data and SVC-RRFE could be a timeless strategy for feature elimination. While most astonishing models made from related studies got high accuracy with extra neuropsychological information, this thesis managed to classify MCI and SCI individuals with text only. The finding of phonemic impairment and patterns in SCI and MCI individuals presented the potential of improving MCI and SCI individual identification. Although 76% accuracy is not good enough for practice, there is always more information to combine to improve the prediction, such as considering other predictors and adding more data points since the SD of the accuracy of MCI group vs SCI group model was slightly high (8.9%) and the ROC AUC could reach 0.84 for certain fold.

Limitations and future directions

A major limitation of the present study is the discourse setting for the speech transcription. While the verb production of the picture description task is most likely present tense than past or feature tense, it provides less POS information to extract compare to story-telling or free conversation (Bhatia, 1993). With respect to the goal of classifying individuals with MCI and SCI, the small size of data (using speech transcription of 23 SCI patients, 31 MCI patients, and 36 healthy controls) limited the performance of machine learning models with less precision and more uncertainty. Regarding the linguistic features explored in the thesis, improvements can be addressed by extending the POS, syntactic complexity and phonology category. Such as, function words, content words, ratio of POS tags corresponding the explored noun-verb ratio; context free production rules, parse tree height, depth of syntactic trees; trigram PSF, long-short vowel ratio, unrounded-rounded vowel ratio, front-back vowel ratio, frequency of consonants, etc. As the process of model building was a proof of concept, a code refinement can be done to make the SVM-RRFE algorithm reusable for further development. To test the availability of SVM-RRFE, other dataset could be introduced to verify and compare with the other feature engineering techniques, such as SVD. Since related studies gained promising outcomes from automatic speech recognition (ASR) in other languages, we can try applying the technology on the current data to improve the automation of the speech data analysis. Moreover, we can gather more unseen data to validate the model comprehensively and make predictions through concatenated models from the k-fold neural networks.

In conclusion, this thesis work showed the importance of phonological patterns on identifying MCI and SCI subjects and provided a complete methodology for the classification tasks at linguistic aspects with novel feature selection strategies and neural networks. It contributes to the community of early diagnosis of AD and other neurological conditions through computational approaches.

Reference

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., … Zheng, X. (2016). *TensorFlow: A System for Large-Scale Machine Learning*. 265–283. https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi

Aguiar, A. C. P. de O., Ribeiro, M. I., & Jacinto, A. F. (2010). Subjective memory complaints in the elderly may be related to factors other than cognitive deficit. *Dementia & Neuropsychologia*, *4*(1), 54–57. https://doi.org/10.1590/S1980-57642010DN40100009

Alberdi, A., Aztiria, A., & Basarab, A. (2016). On the early diagnosis of Alzheimer’s Disease from multimodal signals: A survey. *Artificial Intelligence in Medicine*, *71*, 1–29. https://doi.org/10.1016/j.artmed.2016.06.003

Alberti, C., Andor, D., Bogatyy, I., Collins, M., Gillick, D., Kong, L., Koo, T., Ma, J., Omernick, M., Petrov, S., Thanapirom, C., Tung, Z., & Weiss, D. (2017). *SyntaxNet Models for the CoNLL 2017 Shared Task*. 6.

Alegria, R., Gallo, C., Bolso, M., dos Santos, B., Prisco, C. R., Bottino, C., & Ines, N. M. (2013). Comparative study of the uses of grammatical categories: Adjectives, adverbs, pronouns, interjections, conjunctions and prepositions in patients with Alzheimer’s disease. *Alzheimer’s & Dementia*, *9*(4, Supplement), P882. https://doi.org/10.1016/j.jalz.2013.08.233

Asgari, M., Kaye, J., & Dodge, H. (2017). Predicting mild cognitive impairment from spontaneous spoken utterances. *Alzheimer’s & Dementia: Translational Research & Clinical Interventions*, *3*(2), 219–228. https://doi.org/10.1016/j.trci.2017.01.006

Balash, Y., Mordechovich, M., Shabtai, H., Giladi, N., Gurevich, T., & Korczyn, A. D. (2013). Subjective memory complaints in elders: Depression, anxiety, or cognitive decline? *Acta Neurologica Scandinavica*, *127*(5), 344–350. https://doi.org/10.1111/ane.12038

Barnes, D. E., Alexopoulos, G. S., Lopez, O. L., Williamson, J. D., & Yaffe, K. (2006). Depressive Symptoms, Vascular Disease, and Mild Cognitive Impairment: Findings From the Cardiovascular Health Study. *Archives of General Psychiatry*, *63*(3), 273–279. https://doi.org/10.1001/archpsyc.63.3.273

Basilakos, A. (2016). *Towards improving the evaluation of speech production deficits in chronic stroke* [ProQuest Dissertations Publishing]. http://search.proquest.com/docview/1845017933/?pq-origsite=primo

Beland, R. (1999). Phonological Spelling in a Dat Patient: The Role of the Segmentation Subsystem in the Phoneme-to-Grapheme Conversion. *Cognitive Neuropsychology*, *16*(2), 115–155. https://doi.org/10.1080/026432999380924

Beltrami, D., Calzà, L., Gagliardi, G., Ghidoni, E., Marcello, N., Favretti, R. R., & Tamburini, F. (2016). *Automatic Identification of Mild Cognitive Impairment through the Analysis of Italian Spontaneous Speech Productions*. 8.

Berube Shauna, Nonnemacher Jodi, Demsky Cornelia, Glenn Shenly, Saxena Sadhvi, Wright Amy, Tippett Donna C., & Hillis Argye E. (2019). Stealing Cookies in the Twenty-First Century: Measures of Spoken Narrative in Healthy Versus Speakers With Aphasia. *American Journal of Speech-Language Pathology*, *28*(1S), 321–329. https://doi.org/10.1044/2018_AJSLP-17-0131

Bhalla, R. K., Butters, M. A., Becker, J. T., Houck, P. R., Snitz, B. E., Lopez, O. L., Aizenstein, H. J., Raina, K. D., DeKosky, S. T., & Reynolds, C. F. (2009). Patterns of Mild Cognitive Impairment After Treatment of Depression in the Elderly. *The American Journal of Geriatric Psychiatry*, *17*(4), 308–316. https://doi.org/10.1097/JGP.0b013e318190b8d8

Bhatia, V. K. (1993). *Analysing genre: Language use in professional settings*. Longman.

Borin, L., Forsberg, M., Hammarstedt, M., Rosén, D., Schäfer, R., & Schumacher, A. (2016). *Sparv: Språkbanken's corpus annotation pipeline infrastructure*. 4.

Brodaty, H., & Donkin, M. (2009). Family caregivers of people with dementia. *Dialogues in Clinical Neuroscience*, *11*(2), 217–228.

Brucki, S. M. D., & Nitrini, R. (2009). Subjective memory impairment in a rural population with low education in the Amazon rainforest: An exploratory study. *International Psychogeriatrics*, *21*(1), 164–171. https://doi.org/10.1017/S1041610208008065

Budson, A. E., & Solomon, P. R. (2011). *Memory loss: A practical guide for clinicians*. Elsevier Saunders.

CAN, E., & Gülmira, K. (2018). Assessment of Syntactic Complexity in Alzheimer's disease. *In 3rd Eurasian Conference on Language and Social Sciences*, 517.

Charalambos, T., & Bronte, F. (2018). *Acoustic markers of PPA variants using machine learning*. https://www.frontiersin.org/10.3389/conf.fnhum.2018.228.00092/event_abstract

Chollet, F. (2015). *Keras*. https://github.com/keras-team/keras

Clark, D. G., McLaughlin, P. M., Woo, E., Hwang, K., Hurtz, S., Ramirez, L., Eastman, J., Dukes, R.-M., Kapur, P., DeRamus, T. P., & Apostolova, L. G. (2016). Novel verbal fluency scores and structural brain imaging for prediction of cognitive outcome in mild cognitive impairment. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *2*, 113–122. https://doi.org/10.1016/j.dadm.2016.02.001

Croot, K., Hodges, J. R., Xuereb, J., & Patterson, K. (2000). Phonological and Articulatory Impairment in Alzheimer's Disease: A Case Series. *Brain and Language*, *75*(2), 277–309. https://doi.org/10.1006/brln.2000.2357

Dozat, T. (2016). *INCORPORATING NESTEROV MOMENTUM INTO ADAM*. 4.

Engstrand, O. (2004). *Fonetikens grunder*. Studentlitteratur.

Fraser, K. C., Lundholm Fors, K., Eckerström, M., Öhman, F., & Kokkinakis, D. (2019). Predicting MCI Status From Multimodal Language Data Using Cascaded Classifiers. *Frontiers in Aging Neuroscience*, *11*. https://doi.org/10.3389/fnagi.2019.00205

Fraser, K. C., Lundholm Fors, K., & Kokkinakis, D. (2019). Multilingual word embeddings for the assessment of narrative speech in mild cognitive impairment. *Computer Speech & Language*, *53*, 121–139. https://doi.org/10.1016/j.csl.2018.07.005

Fraser, K. C., Lundholm Fors, K., Kokkinakis, D., & Nordlund, A. (2017). An analysis of eye-movements during reading for the detection of mild cognitive impairment. *Proceedings of* *the 2017 Conference on Empirical Methods in Natural Language Processing*, 1016–1026. https://doi.org/10.18653/v1/D17-1107

Fraser, K. C., Meltzer, J. A., Graham, N. L., Leonard, C., Hirst, G., Black, S. E., & Rochon, E. (2014). Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. *Cortex*, *55*, 43–60. https://doi.org/10.1016/j.cortex.2012.12.006

Fraser, K. C., Meltzer, J. A., & Rudzicz, F. (2016). Linguistic Features Identify Alzheimer's Disease in Narrative Speech. *Journal of Alzheimer's Disease*, *49*(2), 407–422. https://doi.org/10.3233/JAD-150520

Fraser, K. C., Rudzicz, F., & Rochon, E. (2013). *Using Text and Acoustic Features to Diagnose Progressive Aphasia and its Subtypes*. 5.

Garcia-Ptacek, S., Cavallin, L., Kåreholt, I., Kramberger, M. G., Winblad, B., Jelic, V., & Eriksdotter, M. (2014). Subjective Cognitive Impairment Subjects in Our Clinical Practice. *Dementia and Geriatric Cognitive Disorders Extra*, *4*(3), 419–430. https://doi.org/10.1159/000366270

Goodglass, H., & Kaplan, E. (1983). *Boston Diagnostic Aphasia Examination*.

Gosztolya, G., Tóth, L., Grósz, T., Vincze, V., Hoffmann, I., Szatlóczki, G., Pákáski, M., & Kálmán, J. (2016). *Detecting Mild Cognitive Impairment from Spontaneous Speech by Correlation-Based Phonetic Feature Selection*. 107–111. https://doi.org/10.21437/Interspeech.2016-384

Gosztolya, G., Vincze, V., Tóth, L., Pákáski, M., Kálmán, J., & 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. https://doi.org/10.1016/j.csl.2018.07.007

Grinberg, M. (2014). *Flask Web Development: Developing Web Applications with Python* (1st ed.). O'Reilly Media, Inc.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, *46*(1), 389–422. https://doi.org/10.1023/A:1012487302797

Hansson, K., Nettelbladt, U., & Leonard, L. B. (2000). Specific Language Impairment in Swedish: The Status of Verb Morphology and Word Order. *Journal of Speech, Language, and Hearing Research*, *43*(4), 848–864. https://doi.org/10.1044/jslhr.4304.848

Hernández-Domínguez, L., Ratté, S., Sierra-Martínez, G., & Roche-Bergua, A. (2018). Computer-based evaluation of Alzheimer's disease and mild cognitive impairment patients during a picture description task. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *10*, 260–268. https://doi.org/10.1016/j.dadm.2018.02.004

Jarrold, W., Peintner, B., Wilkins, D., Vergryi, D., Richey, C., Gorno-Tempini, M. L., & Ogar, J. (2014). Aided diagnosis of dementia type through computer-based analysis of spontaneous speech. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 27–37. https://doi.org/10.3115/v1/W14-3204

Jesdoit/parsey-universal-server. (2018). [Python]. jesdoit. https://github.com/jesdoit/parsey-universal-server (Original work published 2018)

John, G. H., Kohavi, R., & Pfleger, K. (1994). Irrelevant Features and the Subset Selection Problem. In W. W. Cohen & H. Hirsh (Eds.), *Machine Learning Proceedings 1994* (pp. 121–129). Morgan Kaufmann. https://doi.org/10.1016/B978-1-55860-335-6.50023-4

Kathleen C. Fraser, Fors, K. L., Eckerström, M., Themistocleous, C., & Kokkinakis, D. (2018). Improving the Sensitivity and Specificity of MCI Screening with Linguistic Information. *LREC Workshop: RaPID-2. Miyazaki, Japan*.

Kave, G., & Levy, Y. (2003). Morphology in picture descriptions provided by persons with Alzheimer's disease. *Journal of Speech, Language, and Hearing Research; Rockville*, *46*(2), 341–352.

Kokkinakis, D., Lundholm Fors, K., Björkner, E., & Nordlund, A. (2017). Data Collection from Persons with Mild Forms of Cognitive Impairment and Healthy Controls—Infrastructure for Classification and Prediction of Dementia. *Proceedings of the 21st Nordic*

Conference on Computational Linguistics, 172–182. https://www.aclweb.org/anthology/W17-0220

Koppara, A., Wagner, M., Lange, C., Ernst, A., Wiese, B., König, H.-H., Brettschneider, C., Riedel-Heller, S., Luppa, M., Weyerer, S., Werle, J., Bickel, H., Mösch, E., Pentzek, M., Fuchs, A., Wolfsgruber, S., Beauducel, A., Scherer, M., Maier, W., & Jessen, F. (2015). Cognitive performance before and after the onset of subjective cognitive decline in old age. *Alzheimer's & Dementia : Diagnosis, Assessment & Disease Monitoring*, *1*(2), 194–205. https://doi.org/10.1016/j.dadm.2015.02.005

Kryscio, R. J., Abner, E. L., Cooper, G. E., Fardo, D. W., Jicha, G. A., Nelson, P. T., Smith, C. D., Van Eldik, L. J., Wan, L., & Schmitt, F. A. (2014). Self-reported memory complaints: Implications from a longitudinal cohort with autopsies. *Neurology*, *83*(15), 1359–1365. https://doi.org/10.1212/WNL.0000000000000856

Linz, N., Lundholm Fors, K., Lindsay, H., Eckerström, M., Alexandersson, J., & Kokkinakis, D. (2019). Temporal Analysis of the Semantic Verbal Fluency Task in Persons with Subjective and Mild Cognitive Impairment. *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, 103–113. https://doi.org/10.18653/v1/W19-3012

Lundeborg, I., Nordin, E., Zeipel-Stjerna, M., & McAllister, A. (2015). Voice onset time in Swedish children with phonological impairment. *Logopedics Phoniatrics Vocology*, *40*(4), 149–155. https://doi.org/10.3109/14015439.2014.934276

Lundholm Fors, K., Fraser, K., & Kokkinakis, D. (2018). Automated Syntactic Analysis of Language Abilities in Persons with Mild and Subjective Cognitive Impairment. *Studies in Health Technology and Informatics*, *247*, 705–709.

Macoir, J., Lafay, A., & Hudon, C. (2019). Reduced Lexical Access to Verbs in Individuals With Subjective Cognitive Decline. *American Journal of Alzheimer's Disease & Other Dementiasr*, *34*(1), 5–15. https://doi.org/10.1177/1533317518790541

McDougall, G. J., Becker, H., & Arheart, K. L. (2006). Older Adults in the SeniorWISE Study At Risk for Mild Cognitive Impairment. *Archives of Psychiatric Nursing*, *20*(3), 126–134. https://doi.org/10.1016/j.apnu.2005.09.003

McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*, 56–61. https://doi.org/10.25080/Majora-92bf1922-00a

Meilán, J. J. G., Martínez-Sánchez, F., Carro, J., Sánchez, J. A., & Pérez, E. (2012). Acoustic Markers Associated with Impairment in Language Processing in Alzheimer's Disease. *The Spanish Journal of Psychology*, *15*(2), 487–494. https://doi.org/10.5209/rev_SJOP.2012.v15.n2.38859

Mortensen, D. R., Dalmia, S., & Littell, P. (2018, May). Epitran: Precision G2P for Many Languages. *Proceedings of the 11th Language Resources and Evaluation Conference*. https://www.aclweb.org/anthology/L18-1429

Orimaye, Sylvester O., Wong, J. S.-M., Golden, K. J., Wong, C. P., & Soyiri, I. N. (2017). Predicting probable Alzheimer's disease using linguistic deficits and biomarkers. *BMC Bioinformatics*, *18*(1), 34. https://doi.org/10.1186/s12859-016-1456-0

Orimaye, Sylvester Olubolu, Wong, J. S.-M., & Golden, K. J. (2014). *Learning Predictive Linguistic Features for Alzheimer's Disease and related Dementias using Verbal Utterances*. 78–87. https://doi.org/10.3115/v1/W14-3210

Pakhomov, S. V. S., Smith, G. E., Chacon, D., Feliciano, Y., Graff-Radford, N., Caselli, R., & Knopman, D. S. (2010). Computerized Analysis of Speech and Language to Identify Psycholinguistic Correlates of Frontotemporal Lobar Degeneration: *Cognitive and Behavioral Neurology*, *23*(3), 165–177. https://doi.org/10.1097/WNN.0b013e3181c5dde3

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. (2012). Scikit-learn: Machine Learning in Python. *MACHINE LEARNING IN PYTHON*, 6.

Petersen, R. C., Smith, G. E., Waring, S. C., Ivnik, R. J., Tangalos, E. G., & Kokmen, E. (1999). Mild Cognitive Impairment: Clinical Characterization and Outcome. *Archives of Neurology*, *56*(3), 303–308. https://doi.org/10.1001/archneur.56.3.303

Rabin, L. A., Smart, C. M., & Amariglio, R. E. (2017). Subjective Cognitive Decline in Preclinical Alzheimer's Disease. *Annual Review of Clinical Psychology*, *13*(1), 369–396. https://doi.org/10.1146/annurev-clinpsy-032816-045136

Reisberg, B., & Gauthier, S. (2008). Current evidence for subjective cognitive impairment (SCI) as the pre-mild cognitive impairment (MCI) stage of subsequently manifest Alzheimer's disease. *International Psychogeriatrics*, *20*(1), 1–16. https://doi.org/10.1017/S1041610207006412

Roark, B., Mitchell, M., Hosom, J.-P., Hollingshead, K., & Kaye, J. (2011). Spoken Language Derived Measures for Detecting Mild Cognitive Impairment. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(7), 2081–2090. https://doi.org/10.1109/TASL.2011.2112351

Rye, A. S. (2019). *Andersrye/parsey-universal-server* [Python]. https://github.com/andersrye/parsey-universal-server (Original work published 2016)

Santos, L. B. dos, Corrêa Jr, E. A., Oliveira Jr, O. N., Amancio, D. R., Mansur, L. L., & Aluísio, S. M. (2017). Enriching Complex Networks with Word Embeddings for Detecting Mild Cognitive Impairment from Speech Transcripts. *ArXiv:1704.08088 [Cs]*. http://arxiv.org/abs/1704.08088

Satt, A., Sorin, A., Toledo-Ronen, O., Barkan, O., Kompatsiaris, Y., Kokonozi, A., & Tsolaki, M. (2013). Evaluation of speech-based protocol for detection of early-stage dementia. *INTERSPEECH*.

Studart, A., & Nitrini, R. (2016). Subjective cognitive decline: The first clinical manifestation of Alzheimer's disease? *Dementia & Neuropsychologia*, *10*(3), 170–177. https://doi.org/10.1590/S1980-5764-2016DN1003002

Szatloczki, G., Hoffmann, I., Vincze, V., Kalman, J., & Pakaski, M. (2015). Speaking in Alzheimer's Disease, is That an Early Sign? Importance of Changes in Language

Abilities in Alzheimer’s Disease. *Frontiers in Aging Neuroscience*, *7*. https://doi.org/10.3389/fnagi.2015.00195

Themistocleous, C., Eckerström, M., & Kokkinakis, D. (2018). Identification of Mild Cognitive Impairment From Speech in Swedish Using Deep Sequential Neural Networks. *Frontiers in Neurology*, *9*. https://doi.org/10.3389/fneur.2018.00975

Themistocleous, C., Ficek, B., Webster, K., Ouden, D.-B. den, Hillis, A. E., & Tsapkini, K. (2020). Automatic subtyping of individuals with Primary Progressive Aphasia. *BioRxiv*, 2020.04.04.025593. https://doi.org/10.1101/2020.04.04.025593

Themistocleous, C., Kokkinakis, D., Eckerström, M., Fraser, K., & Fors, K. L. (2018). *Effects of Mild Cognitive Impairment on vowel duration*. 4.

Tóth, L., Gosztolya, G., Vincze, V., Hoffmann, I., Szatlóczki, G., Biró, E., Zsura, F., Pákáski, M., & Kálmán, J. (2015). Automatic detection of Mild cognitive impairment from spontaneous speech using ASR. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2694–2698. https://hungary.pure.elsevier.com/hu/publications/automatic-detection-of-mild-cognitive-impairment-from-spontaneous

Tóth, L., Hoffmann, I., Gosztolya, G., Vincze, V., Szatlóczki, G., Bánréti, Z., Pákáski, M., & Kálmán, J. (2018). A Speech Recognition-based Solution for the Automatic Detection of Mild Cognitive Impairment from Spontaneous Speech. *Current Alzheimer Research*, *15*(2), 130–138. https://doi.org/10.2174/1567205014666171121114930

Vincze, V., Gosztolya, G., Tóth, L., Hoffmann, I., Szatlóczki, G., Bánréti, Z., Pákáski, M., & Kálmán, J. (2016). Detecting Mild Cognitive Impairment by Exploiting Linguistic Information from Transcripts. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 181–187. https://doi.org/10.18653/v1/P16-2030

Wallin, A., Nordlund, A., Jonsson, M., Lind, K., Edman, Å., Göthlin, M., Stålhammar, J., Eckerström, M., Kern, S., Börjesson-Hanson, A., Carlsson, M., Olsson, E., Zetterberg, H., Blennow, K., Svensson, J., Öhrfelt, A., Bjerke, M., Rolstad, S., & Eckerström, C.

(2016). The Gothenburg MCI study: Design and distribution of Alzheimer’s disease and subcortical vascular disease diagnoses from baseline to 6-year follow-up. *Journal of Cerebral Blood Flow and Metabolism: Official Journal of the International Society of Cerebral Blood Flow and Metabolism*, *36*(1), 114–131. https://doi.org/10.1038/jcbfm.2015.147

Wei, Q., Franklin, A., Cohen, T., & Xu, H. (2018). Clinical text annotation – what factors are associated with the cost of time? *AMIA Annual Symposium Proceedings*, *2018*, 1552.

Yasuno, F., Kazui, H., Yamamoto, A., Morita, N., Kajimoto, K., Ihara, M., Taguchi, A., Matsuoka, K., Kosaka, J., Tanaka, T., Kudo, T., Takeda, M., Nagatsuka, K., Iida, H., & Kishimoto, T. (2015). Resting-state synchrony between the retrosplenial cortex and anterior medial cortical structures relates to memory complaints in subjective cognitive impairment. *Neurobiology of Aging*, *36*(6), 2145–2152. https://doi.org/10.1016/j.neurobiolaging.2015.03.006

Yates, J. A., Clare, L., Woods, R. T., & CFAS, M. (2017). Subjective memory complaints, mood and MCI: A follow-up study. *Aging & Mental Health*, *21*(3), 313–321. https://doi.org/10.1080/13607863.2015.1081150

Yu, B., Quatieri, T. F., Williamson, J. R., & Mundt, J. C. (2015). Cognitive impairment prediction in the elderly based on vocal biomarkers. *INTERSPEECH*, 3734–3738.