School of Electronic Engineering and Computer Science

MSc. Big Data Science Thesis Project Report 2017

Interactional and Linguistic Analysis for Computationally Diagnosing Alzheimer's Disease

Claire Mary Kelleher
Supervisor: Dr. Matthew Purver

Queen Mary
University of London

Disclaimer

This report, with any accompanying documentation and/or implementation, is submitted as part requirement for the degree of MSc. in Big Data Science at Queen Mary University of London. It is the product of my own labour except where indicated in the text. The report may be freely copied and distributed provided the source is acknowledged.

Table of Contents

Abstract

Background: On top of memory loss and linguistic impairment, changes in behaviour and decreased interactional skills in conversation are also symptoms for Alzheimer's Disease (AD) and other dementias. Relatively few studies have used computational techniques to investigate the diagnosis predicting power of this layer of symptoms. This project hypothesizes that by adding this new layer symptoms (interactional features) into a classifier used to diagnose AD types, it will increase overall performance.

Method: Using data derived from the DementiaBank corpus [1], this project looks at how a change in a person's interactional features in conversation can be used to better classify the different diagnosis of AD and other dementias using NLP methods and other computational techniques. Using machine learning feature selection methods and the same data from DementiaBank, Fraser et al. [2] have found the optimal number (35) and strongest predicting features that are linguistic, information content based (non-interactional), that best classify AD. This project encoded these non-interactional features and ran them through the same classifier as Fraser et al. [2]. This outputted a similar classifier accuracy score to that of Fraser et al. [2] and so allowing this score to be used as an evaluation baseline to compare how the classifier performs when other features were inputted. Interactional features that are known symptoms of AD and dementia [3] and are used in Svennevig and Lind [4], such as turn-taking, filler term frequency and trailing-off mid sentence, were then encoded. The two groupings of features (non-interactional and interactional) were then compared by first, ranking the features based on classifying weight and secondly, running the top predicting features from this set through the classifier built earlier and comparing the change in accuracy. Correlation analysis was carried out on the interactional features as a sense check and to investigate the direction in which each variable correlates with the diagnosis.

Results: The final top predicting features were made up of 52.2% interactional and 47.8% non-interactional with classifier accuracy scores seeing an improvement by 4.96%.

Conclusion: Since interactional features replaced over half of the originally all non-interactional top predicting features regarding diagnosis predicting weight, as well as causing an improvement in the classifier's accuracy, one can conclude that encoding interactional features, in particular dialogue based features that represent the amount of involvement the invigilator is required to manifest throughout the session, in addition to non-interactional features, can assist in computationally classifying Alzheimer's disease.

Background

Dementia & Alzheimer’s Disease

The word ‘dementia’ describes a set of symptoms that may include memory loss and difficulties with thinking, problem-solving or language. Dementia is caused when the brain is damaged by diseases, such as Alzheimer’s disease or a series of strokes. AD is the most common cause of dementia, but not the only one. The specific symptoms that someone with dementia experiences will depend on the parts of the brain that are damaged and the disease that is causing the dementia.[3] Dementia is a terminal condition and affects millions of people worldwide. Symptoms of dementia include memory loss, confusion and problems with speech and understanding. Due to Dementia having a high prevalence and level of associated morbidity, it has become an urgent health and economic issue for the developed world, and a rapidly growing threat in developing countries. There are currently 850,000 people with dementia in the UK, with numbers set to rise to over 1 million by 2025. Due to the general rise of elderly individuals in the population, this will soar to 2 million by 2051. [5]

AD is a degenerative brain disorder and the most common type of dementia, affecting 62% of those diagnosed. Other types of dementia include; vascular dementia, affecting 17% of those diagnosed, mixed dementia affecting 10% [5]. The risk of developing AD becomes greater with age. AD can refer to a set of symptoms in different cognitive and linguistic domains, and characteristically, these symptoms are persistent and progressive, causing a deterioration of skills and knowledge. The domains affected are memory, executive functions, language, visual-spatial processing, personality and general behavior and interaction skills [4]. This project focuses on exploring the effects that AD has on both the impairment of linguistic and interaction skills using computational techniques.

Project Introduction

After memory loss, language impairment and changes in a patient's behaviour are part of the top symptoms of AD. A number of studies have already been carried out using computational techniques on linguistic features in order to identify Alzheimer’s Disease [2,6,7,8], however relatively few studies have been carried out using computational techniques on features that encodes the patient's interactional behaviour in conversation. It is hypothesized that by including this level of symptoms that has not been investigated before in the AD classification process, the performance of the classifier will improve.

The aim of this project was to encode a set of interactional features (IF) that represent common interactional behaviour changes in AD patients such as difficulties following a conversation or finding the right word for something [3], hesitation, restlessness [9] (eg. conversation restarts, filler terms, pauses) and to investigate the predictive power of these features in classifying the different AD severities both on their own and combined with the non-interactional features that are already proven to be useful [2] in making these predictions. An investigation into the correlation between the interactional features and the diagnosis variables was also carried out in order to investigate in which features most correlate with it.

Literature Review

In Fraser et al.[2], “Linguistic Features Identify Alzheimer's Disease in Narrative Speech”, Fraser, Meltzer and Rudzicz demonstrated state-of-the-art accuracy in automatically identifying AD from short narrative samples elicited with a picture description task, and to uncover the salient linguistic factors with a statistical factor analysis. Data was derived from the same DementiaBank corpus that this project used, from which 167 patients diagnosed with “possible” or “probable” AD provide 240 narrative samples, and 97 controls provide an additional 233 files from 97 speakers. Fraser et al. obtained classification accuracies of over 81% in distinguishing individuals with AD from those without based on short samples of their language on the Boston ‘Cookie-Theft’ picture description task.

This task instructs the examiner to show the picture, which if of a scene in a kitchen where a boy is robbing cookies, to the patient and say, “Tell me everything you see going on in this picture.” The examiner is permitted to encourage the patient to keep going if they do not produce very many words. This leads to the interactions between the participant and the investigator that can be investigated for this project.

Verbal picture description is one of the most sensitive tests for detecting language disorders in early AD [10] and is the reason this task is appears to be used in a number of studies relating to dementia. [6,11,12] Fraser et al. also investigated the heterogeneity of features among the participants of the study. Using factor analysis, four clear factors emerged: semantic impairment, acoustic abnormality, syntactic impairment, and information impairment. A large number (370) of features were considered to capture a wide range of linguistic phenomena. The results and feature groupings from this study played a significant role in the non-interaction feature design for this project. The papers classifier accuracy and build was also used as a baseline for the evaluation when comparing both the interactional and non-interactional features.

Croisile et al. 1996 [6], “Comparative Study of Oral and Written Picture Description in Patients with Alzheimer’s Disease” uses the same ‘Cookie-Theft’ picture task as In Fraser et al.[2] and this project on its participants. These participant consisted of 22 patients with AD and 24 healthy elderly subjects The purpose of the study was to provide comparative information about lexical, syntactic, and semantic aspects of oral and written picture descriptions in AD patients and healthy elderly subjects. They analyzed the similarities and differences of oral and written descriptions by comparison of the results obtained in each group and identified specific impaired features of description processing in AD patients by making an intergroup comparison of the results obtained for each task. The result was that AD patients had a significant reduction of all word categories, which, similarly to controls, was more pronounced in written than in oral texts and in sum, AD descriptions were always shorter and less informative than control texts.

This paper's method to building an “information unit” was used to build certain variables for the classifier. An information unit was used in this study as a means to measure the information content that the participant described. The list consisted of 23 information units in four key categories: subjects, places, objects, and actions. For example, the three subjects were: the boy, the girl, and the woman. If a participant mentioned mother or female, this would count as a mark for the woman information unit.

Svennevig and Lind 2016 [4], “Dementia, interaction, and bilingualism: An exploratory case study”, looks at Norwegian speaking elderly persons with dementia who are multilingual. The study presents an exploratory, clinical linguistic case study of one bilingual speaker diagnosed with probable dementia of the Alzheimer type in two conversational contexts, English and Norwegian. The study explores his speech production in the two languages, focusing on case where the participant displays problems of achieving progressivity of talk and his methods in which he searches for ways of continuing his turn at talk. The study investigates turn-taking and take the most crucial word of an utterance to be the lexical verb as so much of the semantic and syntactic structure of the utterance, hence also the interpretation of the utterance, depends on the choice of the lexical verb. Without lexical verbs, utterance interpretation is very challenging, even in context. The study look at the participant's word-finding difficulties manifested as a lack of progressivity of the talk and investigate the way in which the participant solves these difficulties. For example, solutions for this difficulty would include the speaker saying semantically meaningless nouns (eg. “thing”) as a substitute for the intended word or fillers (eg. “uh”, “em”). Sometimes, the act of searching is not solely caused by the speaker but a joint effort. The speaker may include an invitation to the other party in the search, for instance by gazing at him or her, or by explicitly appealing for assistance. The data collected included responses to formal cognitive and linguistic tests, as well as responses to a questionnaire on functional communication and recordings of more or less spontaneously produced speech (elicited narratives and conversation). The article focuses primarily on the conversational data, while using some of the other data as background data for the description of the participant.

Although the above exploratory study focusing on multi-linguists in Norway, the interactional features mentioned above, plus a number more used in the paper can be said to measure anomia, a symptom of AD. This is the reason behind a number of features were created based around the turn-taking and tokens that may represent interruptions in an utterance and difficulties in progressivity of talk in the hope that they would be strong predictors of the diagnosis.

In [13] Aphasia therapy dialogues, Silvast, M. Investigates the interaction between aphasic patients and the speech therapist in order to investigate the role of the therapist when using conversation as a method for rehabilitating aphasic patients. Aphasia, is the inability to understand or produce speech and has been a proven symptom of AD [14]. This interactions between the patient and the therapist were explored by videotaping a fraction form a therapy session. Six aphasic-therapist pairs served as subjects in the study. A middle five-minute segment of each conversation was extracted for analysis, which focused on the use of interactional space and different communicative functions in therapy conversations. The results showed that during the conversation, therapists had a regulatory role which was manifested in their frequent use of requests for information and clarification. Variables such as TurnCountRatio and ExaminerQuestionCount are built in an attempt to represent and investigate this phenomenon as this will measure the presence of the examiner throughout the session. Aphasics had more speech time but were in a responsive role which again, can be encoded by the question frequency of the examiner. The reason the aphasiacs had a longer speech time is explained by their frequent trouble-indicating behaviour during speaking such as pauses, fillers and repeats. This project investigates different aspects of these three phenomena.

Motivation for investigating Interactional Features

There has been a number of studies carried out using computational techniques on the linguistic based symptoms of dementia [2][6] however there has been relatively little research done on symptoms that fall under what can be classified as interactional symptoms such how a patient solve anomia (word retrieval problems) and what solution path they take. As mentioned in the literature review, individuals with dementia of the Alzheimer’s type often experience anomia [4] leading to the use of filler terms. Hesitant speech also increases with the severity level of the dementia [15], again a known situation where the speaker either intentionally or unintentionally utters filler terms or pauses. Other non-linguistic, interactional features that become more prominent as AD progresses include the decline in the patient's ability to concentrate, making it difficult for the patient to complete tasks and to follow conversation. [3] This can phenomenon can be encoded by analyzing the length of an AD patient’s answers or the if they needed to clarify the task at hand by asking the examiner questions, which has been proven to happen with aphasic patients [13].

By investigating theses interactional features, another layer of dementia symptoms can be added to the analysis which have not been looked at using computational techniques and NLP methods before. It is theorized that by adding in more variables that represent another layer of dementia symptoms, the diagnoses of dementia types using computational techniques should have a higher accuracy result.

Hypothesis

It is hypothesised that the addition of the interactional features mentioned above will improve the classification of AD that are only using with Non-IF’s. This is because the additional interactional features represent another level of symptoms used to diagnose AD as opposed to only taking into account the patient's linguistic symptoms.

Materials & Method

Dataset

The data are derived from the DementiaBank (Pitt) corpus [1] downloaded in June 2017, which is part of the larger TalkBank project. [16] These data were collected between 1983 and 1988 as part of the Alzheimer Research Program at the University of Pittsburgh. [17] Participants were referred directly from the Benedum Geriatric Center at the University of Pittsburgh Medical Center, and others were recruited through the Allegheny County Medical Society, local neurologists and psychiatrists, and public service messages on local media. To be eligible for inclusion in the study, all participants were required to be above 44 years of age, have at least 7 years of education, have no history of nervous system disorders or be taking neuroleptic medication, have an initial Mini-Mental State Exam (MMSE) score of 10 or greater, and be able to give informed consent. Additionally, participants with dementia were required to have a relative or caregiver to act as an informant. All participants received an extensive neuropsychological and physical assessment. Participants were assigned to the “patient” group primarily based on a history of cognitive and functional decline, and the results of a mental status examination. In 1992, several years after the study had ended, the final diagnosis of each patient was reviewed on the basis of their clinical record and any additional relevant information (in some cases, autopsy).

The Pitt corpus that was used for this project consisted of 550 transcripts. Each transcript is a recording between one participant and the examiner where the examiner used the “Cookie Theft” picture description task from the Boston Diagnostic Aphasia Examination [18] on the participant. This test consists of the examiner telling the participant to describe everything that they see in the picture that is being shown to them. The pictures is a scene set in a kitchen, where a boy is robbing cookies while his mother is cleaning the dishes. If the participant does not produce many words, the examiner is allowed to encourage the patient to continue to try and produce some more. Each transcript is one session, and so represents one patient with one diagnosis. Variables for this analysis were created based on one value (average or otherwise) that represents the feature in that particular transcript. In other words, the data that the model was trained and tested on was at transcript level and each features was represented by a single number for each transcript. As mentioned earlier, their was one target value per transcript so this was already on the correct level (transcript) for analysis. The five diagnoses (target) labels were ‘MCI’ (Mild Cognitive Impairment), ‘Memory’, ‘Possible AD’, ‘Probable AD’, ‘Vascular’ (Vascular Dementia), where these approximately in order of severity respectively.

MCI is an intermediate stage between the expected cognitive decline of normal aging and the more-serious decline of dementia. It can involve problems with memory, language, thinking and judgment that are greater than normal age-related changes. [19] There were 43 participants in the study with MCI. ‘Memory’ classifies patients that are experiencing memory loss only and at a rate only slightly greater than that of the changes that would be expected for a person at that age. The data only consisted of 3 ‘Memory’ diagnosis. Possible AD and Probable AD are the different severities of Alzheimer's disease with 'Probable' being more severe than 'Possible'. There was a total of 237 'Probable AD' transcripts and there was only 21 transcripts that recorded patients that were classified with the 'Probable AD' diagnosis. Vascular dementia is caused when blood flow to the brain is reduced and is the second most common cause of dementia, after AD. [20] There were only 5 patients in the DementiaBank data with Vascular dementia. Due to the dataset being unbalanced, precautions were taken in ensure results could be trusted. This included the use of cross-validation when training the model as well as taking into account the recall score (and F1 score) when looking at classifier accuracy scores. Different cuts of the data (eg. 'Control' against 'Probably AD' only) and different groupings of the label's (eg. group any non-AD label as 'Other') were also investigated.

Each speech sample was recorded then manually transcribed at the word level following the TalkBank CHAT (Codes for the Human Analysis of Transcripts) protocol [21]. Narratives were segmented into utterances and annotated with filled pauses, paraphasias, and unintelligible words. CHAT is the standard transcription system for the TalkBank and CHILDES (Child Language Data Exchange System) Projects. All of the transcripts in the TalkBank databases are in CHAT format. These annotations were used to encode and analyse the interactional features of the patients in this study whereas they were not used for the non-interactional (linguistic). Instead, the transcripts were passed through the Stanford Tagger[1] and Stanford Parser[2]. This is because Fraser et al.[2], the paper that the

[1] Version 2017-06-09 https://nlp.stanford.edu/software/tagger.shtml

[2] Version 2017-06-09 https://nlp.stanford.edu/software/lex-parser.shtml

project used as a baseline of evaluation regarding feature encoding and model building, did so and by running the transcripts through the same parser (although a more recent version of the Stanford tagger and parser were used for this project[3]) as Fraser et al.[2], this project was allowed compare the results for the non-interactional features to those of Fraser et al.[2]. They did however, use the manually annotated tags from DementiaBank corpus to test the performance of the Stanford Tagger. More information on other preprocessing steps that had to be taken are discussed in the following sections.s

Method

This project investigates how a change in a person's interactional features in conversation can be used to better classify the different diagnosis of AD using NLP methods and other computational techniques. This means that the main requirement for this project was to produce a set of encoded variables that represent both linguistic and interactional phenomena. Once encoded, these variable sets can be compared against each other regarding their classifying weight towards predicting the target variable. The target variable the five classes mentioned previously of the dementia diagnosis as well as the ‘Control’ group. The five diagnoses labels are ‘MCI’ (Mild Cognitive Impairment), ‘Memory’, ‘Possible AD’, ‘Probable AD’, ‘Vascular’ (Vascular Dementia), where these approximately in order of severity respectively.

Using machine learning feature selection methods and the same data from DementiaBank, Fraser et al. [2] have found the optimal number (35) and strongest

[3] [2] tagger: Version 2015-01-29, [2] parser: Version 2010-11-30

predicting features that are linguistic, information content based (non-interactional), that best classify AD. In order to carry out this analysis, a baseline accuracy score was needed to evaluate the effects of the new feature combinations on classifying the different types of AD. It was decided this paper would be used as this baseline. This is due to it's significant exploration into the what features, the optimal number of features and what combination of features best classify the different severities of AD. The study investigates a large number (370) of features where machine learning and feature selection methods. The maximum highest accuracy score (81%) for their model (logistic regression) was received when 35 features were inputted. This was followed up with an exploratory factor analysis. This analysis was carried out on the top 50 of the 370 features. This number of variables were chosen because it was found that there was not much change in the classifiers accuracy score until more than 50 features were added, where there was then a sharp drop in accuracy score. The factor analysis resulted in finding four factors or groupings of of features: semantic impairment, acoustic abnormality, syntactic impairment, information impairment.

For this project, it was decided to re-build the features from the semantic, syntactic and information impairment factors (in other words, ignoring the acoustic factor) from Fraser et al. [2] and use them to define the Non-Interactional Features (Non-IF's). A total of 29 Non-IF's were encoded. Details on each feature and the reasoning behind the algorithm used to encode each one can be found in the following section. A number of classifiers trained on the DementiaBank data using the above Non-IF's as variables and the, including Decision Tree, Logistic Regression, K-Nearest Neighbour and Naive Bayes.

Different labellings of the target variable were also used. For example, one cut of the data counted any non-AD targets as ‘Other’.

This project encoded these non-interactional features and trained and tested a classifier to replicate that of Fraser et al. [2]. This outputted a similar classifier accuracy score to that of Fraser et al. [2] and so allowing this score to be used as an evaluation baseline to compare how the classifier performs when other features were inputted.

Interactional features that are known symptoms of AD and dementia [3] and are used in Svennevig and Lind [4], such as pause and filler term frequency and trailing-off mid sentence, were then encoded. On top of encoding interaction features based on Svennevig and Lind [4], other IF’s were chosen in a more investigative approach since very little computational analysis has been carried out on these type of features to date. For example, turn-taking ratio between the examiner and the participant was investigated as it is said in [13] that patients that suffer from aphasia tend to require the examiner to have a stronger presence that encourages the participant for more information.

The two groupings of features (non-interactional and interactional) were then compared by first, ranking the features based on classifying weight and secondly, running the top predicting features from this set through the classifier built earlier and comparing the change in accuracy. Correlation analysis was carried out on the interactional features as a sense check and to investigate the direction in which each variable correlates with the diagnosis.

Implementation

TalkBank Manual: CHAT & CLAN

The first step of implementation was to investigate the TalkBank transcripts and have an understanding of what the data represent. In order to this, an understanding of the TalkBank manual was required. There are three parts to the overall TalkBank manual. Part 1 describes the CHAT transcription system, part 2 describes the CLAN (Computerized Language Analysis) analysis programs and part 3 describes the segments of the CLAN program that perform automatic morphosyntactic analysis.

Python

Python (3.6) programming language was used to encode the variables. Python has a range of libraries that proved incredibly useful for this project. The libraries that were utilised included ‘pandas’, ‘numpy’, ‘sklearn’, ‘scipy’ and others. NLTK, Python platform for working with human language data was also utilised.

PreProcessing

The pre-processing step for encoding the Non-IF’s differed from that of the IF’s. This is because when Fraser et al. [2] encoded their features, they only kept the word level transcription and the utterance segmentation. In other words, they discarded the morphological analysis, disfluency annotations, and other associated information that TalkBank had annotated. Whereas, variables created to represent the IF’s needed these annotations as they represent interactional phenomena, both of the participant and between the participant and the invigilator.

For the Non-IF extraction, even after removing their CHAT morphological annotations, there were a still a number of unneeded symbols within the tokens of the transcript. Since the words only were needed for the analysis, these unwanted symbols and digits were removed using regular expression. The tokens were lemmatized and passed through the Stanford Tagger and Stanford Parser (versions stated in earlier ‘Method’ section).

Feature Extraction: Non-Interactional Features

As stated earlier, Fraser et al. [2] have found the optimal number (35) and strongest predicting features that are linguistic and information content based (non-interactional), that best classified AD. Factor analysis was on the top 50 features (out of 370) to finding underlying groupings of the features. The factor analysis resulted in finding four factors or groupings of of features: semantic impairment, acoustic abnormality, syntactic impairment, information impairment. The encoding of the Non-Interactional Feature was based on the features within the non-acoustic factors. Ie. semantic impairment, syntactic impairment, information impairment. Fraser et al. [2] found these factors were found by carrying out a promax oblique rotation. For this project, the features with the most significant factor loading scores (greater than 0.3) were chosen to be encoded to represent the Non-IF’s (see Fraser et al. [2], Table 2). This was so the factor interpretations (groupings) could be carried forward and used throughout the analysis.

The following are the features encoded for this project to represents the Non-IF's grouped by their relative factors that were interpreted by Fraser et al. [2], a brief explanation on how each variable was encoded, including, if needed, how they were normalised and the output. It was not the aim of this project to rebuild the features from Fraser et al. [2] exactly as they did but to use them more as a strong guidance with the aim of building training classifier that resulted in similar accuracies to that of Fraser et al. [2] in order to compare the change in accuracy when the IF's were included. Regarding the POS (Stanford) tags, a dictionary was created for each script that counted the frequency of each POS tag. Each POS count was normalised by the total number of words spoken by the participant.

For the non-interactional features, all variables were encoded at participant level, and the below features of the invigilator were not taken into account for the Non-IF's extraction. So when the following features state "total tokens in transcript", it refers to total tokens in transcript spoken by participant.

Factor 1. Semantic Impairment

Pronoun: Noun Ratio: Count of Pronoun / Count of nouns

Adverbs: Frequency of adverbs / total tokens in transcript.

Verb frequency: Frequency of verbs / total tokens in transcript

Nouns: Frequency of nouns / total tokens in transcript

Word length: Script letter count / script word count. Ie. The average word length per script was used to encode this variable.

Honore’s statistic: Honore’s Statistic attempts a deeper analysis by accounting for words that are only used once, indicating a higher lexical richness therefore a negative Honore’s Statistic suggests low lexical diversity, a known symptom of aphasia and AD. [14] Honore’s Statistic was represented in this project by creating a dictionary that uses a the list of distinct words in the transcript as it’s keys and a frequency count in the transcript of each word. A count of any words that had a frequency of one (ie. Words used once) per transcript was used to represent the Honore’s statistic.

Inflected verbs: Inflection is the name for the extra letter or letters added to nouns, verbs and adjectives in their different grammatical forms. Verbs are inflected in the various tenses (-ing,-s, and-ed) [22]. Therefore, to encode this feature, the different verb inflections were counted. A feature for each inflected verb frequency was created. The five encoded verb inflection variables were past tense, gerund (a verb form which functions as a noun), present participle, non-3rd person singular present, 3rd person singular present. As expected there were was a lot of correlation between these features so some are removed during the feature selection process.

Average cosine distance: This was used to measure participants repetitiveness. Fraser et al. [2] measured the cosine distance between each pair of utterances. To encode this variable for this project, each utterance was represented as a term-document vector. TfidfVectorizer, a function that converts a collection of raw documents (in this case utterance) was used to vectorize the utterances. Once the vectorization and transformation was applied to the pair of utterances that the similarity score was being calculated on, a similarity matrix was produced. This was a 2x2 matrix since it was a one utterance being compared against another utterance, where the entries of the matrix were the cosine similarity score. As the entries of diagonal were all equal to 1 (cosine distance), this was able to be used as a sense check because when the utterance was being compared against itself and it obviously going to have cosine similarity score of 1, as both utterance vectors are in the exact same direction (angle between them is 0).

This process was carried out between every participant's utterance and their utterance prior to that one (with the exception of first utterance as this did not have an utterance before it) to find the cosine distance between each of them. The sum of these cosine distances was calculated. This value was divided by the total number of participants utterances in the session in order to find the value for the ‘Average cosine distance’ variable.

Lexical Parsings: Each utterance of the participant was parsed to find it’s most likely Probability Context Free Grammar (PCFG) tree using the Stanford parser. A variable was created for each of the following parsings was created, which was a count of the occurrence of the parsings throughout the session, divided by the total number of utterances spoken by the participant. It was calculated by traversing through the main parse tree subtrees. If the root of the subtree was equal to to the parent node label (POS tag) in question (before the arrow), then that’s nodes children nodes were traversed through. If these children nodes labels are equal to the POS tag in questions (after the arrow), a value of 1 was added to the counter for this parsing. This total value for this counter was then divided by the total number of utterance spoken by the participant at the end of the script. The following parsings were encoded and counted per participant per session:

ADVP -> RB

NP -> PRP

NP -> DT NN

Factor 2. Syntactic Impairment

Not-In-Dictionary: Count of words that participant mentions that are not in dictionary. A list of words in the NLTK corpus was used as the dictionary for the creation of this variable.

Verbs: Absolute frequency of verbs (base tense)

Verb rate: Frequency of participant verbs / total participant tokens. (As previously stated, this correlates with the other verb features so a number of them were removed in the features selection process further on in the study.)

Lexical Parsings: As explained earlier, all parsing based features was created by parsing the each of the participants utterances using the Stanford parser and analysing each it's subtrees POS labels and counting all the subtrees within the main tree where the parse in question occurred. In order to normalise, this count was then divided by the total number of utterances spoken by the participant. This method was used to create all parsing based features, which is explained in more detail in the following syntactic section. The following parsing occurrences (one variable each) were counted per participant per session:

ROOT -> FRAG

VP -> AUX VP

VP -> VBG

Factor 3. Information Content Impairment

There were three types of features for this factor, keywords, information units and prepositional phrase frequency. The first two types of variables relate directly to the descriptive content stated by the participant (either they have or have not mentioned certain keywords) and the third reflects the detail at which a participant describes this content.

A keyword based features is simply a total count of the times a participant's mentions a keyword. There were four keywords: ‘Window’, ‘Sink’, ‘Cookie’, ‘Curtain’, ‘Counter’. These were encoded into four separate features in order to investigate if anyone particular keyword had a correlation with the target variables, as well as being summed to create another variable that is the total number of keywords mentioned throughout the session.

An information unit is a measure of how much content the participant has described in the description task. Fraser et al. [2] used the following definition of an information unit from Croisile at al.[6] which was used to encode this variable for this project. An information unit was was a list of words categorized into four key categories: subjects, places, objects, and actions. The three subjects were: the boy, the girl, and the woman. The two places were the kitchen and the exterior seen through the window. The eleven objects were: cookie, jar, stool, sink, plate, dishcloth, water, window, cupboard, dishes, and curtains. The seven actions or facts were: boy taking or stealing, boy or stool falling, woman drying or washing dishes/plate, water overflowing or spilling, action performed by the girl, woman unconcerned by the overflowing, woman indifferent to the children. To re-build this feature, utilizing Wordnet's (Python library) 'Synsets', a list of synonyms for each information unit's keyword is created and used to represent an information unit, as opposed to using the exact same list as Croisile et al.[6]. If a participant mention one of the words related to the keywords synonym list, this was a count for the information unit associated with that keyword, as that information unit had been mentioned. The following five keywords were used to build five information units: 'Window', 'Curtain', 'Cookie', 'Sink', 'Girl'.

In order to measure prepositional rates, the POS frequency dictionary that was created earlier was used to find the total number of prepositions stated by the participant, these were normalised by being divided by the total number of words spoken by the participant.

Feature Extraction: Interactional Features

For the purpose of this study, the definition of interactional features represents non-linguistic interactions carried out between the participant and the examiner as well as those of the participant and the examiner separately, such as fillers, pauses and conversational interactions such as turn-taking. There were certain lexical aspect to the creation of these interactional features such as analysing the POS tag that appear directly after the occurrence of a filler.

There were a total of 24 interactional features (see Appendix Table 1 for list) encoded and investigated to determine their diagnosis predicting power (using the same target variable as the non-interactional features). The following section explains the background, reasoning and hypothesis for choosing each variables and how it was encoded. The CHAT annotation was kept and used for the majority of interactional features and analysis of them. The quality of these annotations were sense checked by manually listening to a random selection of 25 transcripts (~5%) and sense checking that the annotations in the transcript were aligned with the recording.

A hierarchical grouping system of the interactional features was created which was based on a number of previous studies showing what interactional symptoms can be used to diagnose AD and other types of dementia.[4][13][23][24][25] There were three hierarchical groupings that are referred to as ‘umbrella features’ where the sub-features within these groupings were different aspects of each umbrella feature. The three umbrella features are fillers, unintentional silence and conversation. The following is list of umbrella features, their sub-features and the reasoning and hypothesis behind choosing them to be investigated for this project.

Umbrella Feature 1. Fillers

A filler is a sound or word that is spoken in conversation by one participant to signal to others a pause to think without giving the impression of having finished speaking. For example, “um”, “like”, “uh”, “you know!” and “actually.” As mentioned in the literature review, individuals with dementia of the Alzheimer’s type often experience, word retrieval problems (anomia) and typically as the disease progresses, the patient’s production of speech decreases (aphasia) and the use of empty phrases, speech automatisms increases, where fillers tend to be used to fill in the lexical gap [4]. This is the reason that fillers were decided to be investigated as interactional features for this project.

A total of 10 different aspects of the filler term were investigated. It is important to note here that the CHAT annotation definition of a filler was used to locate the fillers throughout each session.

Total filler frequency was simply a count of the number of times a participant uttered a filler term. This was normalised by the total number of words spoken by the participant.

POS tags post filler: Five features was encoded to investigate the different POS tags (Adjective, Adverb, Noun, Verb, Other) of the word that follows the filler term. The reason these features were created were to investigate the types of words AD patients were tending to forget. It was hypothesised that participants with AD would be inclined to forget nouns more than any other POS. This hypothesis was constructed because, in [26], where acoustic and POS features were used to distinguish between 9 AD patients and 9 controls, confirmed that AD patients used more pronouns, verbs, and adjectives and fewer nouns.

A table (Appendix Table 5) of higher level POS tags was created so to not include granular POS tags. For example, the POS tag ‘Verb’, included all the different tenses of verb.

A **'Filler Location' score** of each transcript was encoded. This takes into account the location of the filler in the utterance normalised by the total words of the transcript. It is hypothesised that AD patients will tend to forget a word sooner and have to pause and utter fillers earlier on in an utterances compared to non-AD patients. The location of the filler was defined by the index of the filler within the utterance. Therefore, if a filler occurred early in an utterance, it would have a low index value. Ie. If first word of utterance was a filler term, the index of this filler term is equal to zero. For each transcript, all index value were summed and this number was divided by the total number of fillers that were uttered by the participant, this value was a participants 'Filler Location' score. Participants with a low 'Filler Location' score were hypothesised to be diagnosed with AD as they would have uttered a high frequency of fillers (scores denominator) and a low value for the sum of filler term indexes (numerator).

An **average cosine similarity score** (the same method was used to measure repetition of participants in the non-interactional features) of each word that occurred directly after the filler term was calculated for each transcript. If this variable was low this means that the all the words that the participant stated after the filler term were similar. This implies that thy participant was hesitating and forgetting words that were similar to each other. It was hypothesised that this would represent a participant repeating the same mistakes. A high value for this variables would represent a participant that repeated the same mistakes, a known symptom of dementia.[13]

A count of the following fillers were taken, based on the CHAT annotation, all normalised by the total number of words spoken by the participant:

A frequency count of the times a participant **laughed**.

A frequency count of the times a participant **sighed**.

Unintelligible words: A count of unintelligible words were encoded and used to signify the phenomena of mumbling, a common symptom of dementia. [24]

Self-correction counts: A count of the number of times (based on CHAT notations) was also encoded. This variables could be used to represent self-repair, a representative when a person acknowledges that they made a mistake. If this variable negatively correlates with the AD diagnosis, it is hypothesised that this is because the mistake was made however, if it positively correlated with the target, this could be because a non-AD participant has acknowledged their mistake and corrected themselves. This variables fall under the filler umbrella feature because fillers and pauses tend to be used in the process of self-repair between the word that needs repairing and the speaker's intended word.

Umbrella Feature 2. Unintentional Silence

This umbrella feature was used to represent when a moment of silence occurred or an utterance was interrupted or terminated unintentionally. This can happen for a number of reasons such as the participant losing concentration or forgetting what they were supposed to be doing [3]. This has similar reasoning for occurring as the above fillers such as aphasia and word retrieval [13] however, instead of trying to fill the lexical gap with semantically meaningless paraphrase or nouns or other speech automatism, the void in conversation is left as it is.

6 different aspect variables of this umbrella feature were encoded using the CHAT annotations and investigated. These were the different **pause lengths** (long, medium, short) as well as the total seconds that a participant had paused for throughout at the session and a count of when a participant **trailed off mid-utterance** and did not complete the utterance but was also not interrupted.

Incomplete words: The number of incomplete words was also counted, however one the issues with this variables is that it may count a word as unfinished when it was actually the participants accents. Eg. If the participant said "travelin'" instead of "travelling", the programme would flag this as one incomplete word. This was normalised over the total number of words uttered by the participant.

Umbrella Feature 3. Conversation

This umbrella feature was created to represent phenomenon between the participant and the examiner that based on the conversation dominance. Dominance in everyday conversation has been measured by the distribution between speakers of various interactional features, including topic control, interruptions and overlaps, and amount of speech. [23] The reason this area of interaction was investigated for this project was due to [13], an exploratory study carried out between speech therapists and patients of aphasia. It’s results showed that during the conversation therapists had a role which was manifested in their frequent use of requests for information and clarification. Aphasics had more speech time but were in a responsive role in the conversation. Aphasia is a common trait of AD patients, which is shown in [25] where a speech and language assessment in 30 patients with dementia of the Alzheimer type and in 70 normal controls revealed that all AD patients were aphasic. This is linked to the decline in AD patients ability to concentrate, making it difficult for the patient to complete tasks and to follow conversation. [3] Therefore, it is hypothesised in this project, if the participant has AD, the examiner, similar to [13], will need to encourage the participant for more information about the picture throughout their descriptive task. This will result in shorter participant answers to examiner questions, a higher turn-taking rate between the two parties and the examiner having a higher presence throughout the session for participants with AD and other dementias than that of the control.

A total of 7 variables were encoded under this umbrella feature. The following is a list of variables and the algorithm used to encode these interactional phenomenon between the participant and the examiner:

Backchannels: A ratio of the number of times the examiner used a backchannel, such as “Mhm” to signify approval or a suggestion that they are expecting more information from the participant, was calculated by counting the number of backchannels (based on the CHAT annotations) and divided by the number of words spoken by the participant. Therefore, it is suggested that a participant with a high value for this variable would have AD.

Questions: A count of the both the examiner and the participants questions were taken and normalised over their total number of utterances respectively.

Answer Length: The average answer length per transcript was calculated first counting the number number of words the participant stated directly after the examiner's utterances ended in a question mark and up untill the next utterance commenced. This value was then divided by a count of the examiner's question marks in other to find the average answer length per transcript.

Turn Count Ratio: A total count of turns (turnCountINV/turnCountPAR) were also counted within each session. Therefore, the higher the value for this variables, the more utterances the examiner had throughout the session.

The number of times a participant uttered **"I don't know"** or synonyms of that phrase was counted and normalised. This can be used to represent a lack of clarity of the task at hand or misunderstanding between the participant and the examiner.

Non-Interactional Features Selection

A total of 34 non-interactional features were originally encoded (based on Fraser et al. [2]'s non-acoustic top predicting variables) for this project. The top predicting variables, regarding the classification of the diagnosis, were selected from these 34 features. The aimed outcome was a classifier trained on the TalkBank data using these encoded non-interactional features as variables with a resulting similar accuracy score to that of the classifier using similar features and the same data in Fraser et al. [2] to create a baseline to evaluate the interactional features against. In order to do this, it was decided to keep as many variables built as possible, as they are already proven to optimal predictors, and base the feature selection method on the removal of variables that were correlated with each other. This was due to the fact that the phenomena that was being represented were already encoded in some other variables and leaving the overfitting features in can lead to the model overfitting resulting in higher accuracy scores that are not representative of the classifier.

The first step was to investigate the correlation between each of the variables. This was done by creating a correlation matrix (Appendix Fig. 1) with all 34 linguistic features. As anticipated and seen from the correlation matrix below, variables that were built on counts of the letters or words correlated. For example, the average word length of the session (1: 'Avg_Word_Len') correlates with the number of letters within the script (13: 'Script_Letter_Ct') so it was decided to remove the latter. The total count of verbs (23: 'Verb_Ct_Total') naturally correlated with the breakdown of all the different verb tenses (17: 'Verb_Ct_3sing',18: 'Verb_Ct_Base',19: 'Verb_Ct_Gerund',20: 'Verb_Ct_Past',21: 'Verb_Ct_Past_P',22: 'Verb_Ct_Past_T' ,24: 'Verb_Ct_non3sing') as well as the participant's number of ProNouns (11: 'ProNoun_Ct') and the total number of utterances spoken by the participant (9: 'PAR_UTT_ct'). Since it correlated with so many variables it was thought best to, not only run a correlation matrix with it removed and but to also investigate the correlation matrix with the total count of verbs remaining instead of the breakdown of different verbs as well as the different combination of the verb break downs. Any variables where there only difference between them and another variable was the normalisation step, were removed. All the parser based variables remained as they didn't correlated with themselves or any other variables. 11 variables in total were removed in this step so 23 remained. The correlation matrix was built again to confirm there was less correlation between the variables (Appendix Fig. 2). The remaining 23 variables after the feature selection (features in bold font in above legend) were decided to be the top non-interactional features to train the classifier on.

Classifier Build

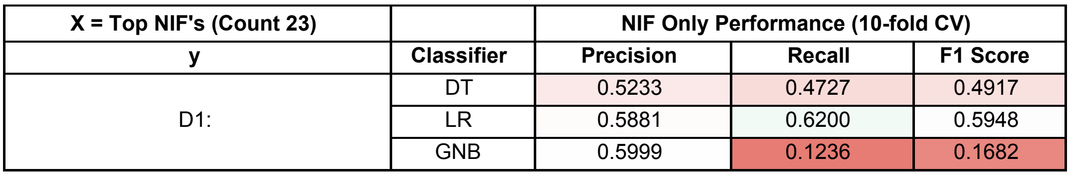
Four different classifiers were trained, each on four different cuts of the data labellings of the target variable. The four classifiers were decision tree, logistic regression, Gaussian Naive Bayes and k-nearest neighbour, which were all built by utilizing the Python library 'sklearn'. The four different cuts of the data and labellings of the data (D1, D2, D3, D4) can be seen in Appendix Table 4.

The decision tree classifier split the data on the best split with it's random state parameter being set to 0. The logistic regression classifier used an L2 regularizer to avoid overfitting. Both the Gaussian Naive Bayes and k-nearest neighbour classifier used the default 'sklearn' parameters.

Due to the relatively small data set of (550 entries, each representing one participant), 10-fold cross validation was used to train and test the data on each classifier in order to avoid overfitting and ensure that the test set was representative of the entire data set. Regarding the target labels, the data was unbalanced; there was a large number of 'Probable AD' and 'Control' participants and a low count of the other targets. To ensure a fair performance metric, precision, recall and F1 scores were all calculated for each fold for each classifier. The average of the 10 folds was then taken as the final performance metric (see Appendix Table 3). The different labelings of the target variables and cuts of the of the data that were mentioned above were also compared.

Fraser et al [2] built their model using logistic regression and the same cut of data, and target labels as ‘Dataset 4’ (D4) in Table 2 below. This merged both severities (‘Possible’ & ‘Probable’) of AD as ‘All AD’ and compared them against the control only. All other dementia causes were ignored. As can be seen from Table 1 below, when the LR classifier was ran on the D4 dataset for this project, the accuracy scores were almost 78%, with both a recall and F1 score of approximately 73% (highlighted in bold in Table 1 below). These accuracy results are similar to that of Fraser et al [2] (~80%) where they applied the same classifier to the same cut of data. It was expected that this project's performance metrics would be slightly lower than those of Fraser et al [2] since the acoustic features were not encoded and so less features were used.

It has been shown in Fraser et al [2] what the optimal non-interactional features are to predict the different diagnosis of AD. This project re-encoded these features (NIF’s) and removed variables that correlated amongst each other. Due to the similarity in performance metrics of the LR classifier for D4 in this project and those of Fraser et al [2] where the same data, target labels and classifier method were used, these remaining variables (Top NIF’s) can be used as a baseline to evaluate the performance of the interactional features.



X = Top NIF's (Count 23)		NIF Only Performance (10-fold CV)		
y	Classifier	Precision	Recall	F1 Score
D1:	DT	0.5233	0.4727	0.4917
	LR	0.5881	0.6200	0.5948
	GNB	0.5999	0.1236	0.1682

	KNN	0.5277	0.5145	0.5046
D2:	DT	0.6197	0.5670	0.5810
	LR	0.6991	0.6831	0.6821
	GNB	0.7587	0.2306	0.3096
	KNN	0.6136	0.5789	0.5838
D3:	DT	0.5160	0.4636	0.4816
	LR	0.5889	0.6200	0.5953
	GNB	0.5983	0.1309	0.1745
	KNN	0.5346	0.5109	0.5044
D4:	DT	0.7117	0.6293	0.6517
	LR	**0.7750**	**0.7233**	**0.7381**
	GNB	0.7488	0.6914	0.7094
	KNN	0.7102	0.6231	0.6455

Table 1. Count of data for the different labellings of the target variable.

	Control	Possible AD	Probable AD	MCI	Vascular	Memory
D1:	241	21	237	43	5	3
D2:	241	21	237	0	0	0
D3:	241	21	237	51 ('Other')		
D4:s	241	258 ('All AD')		0	0	0

Table 2. Count of data for the different labellings of the target variable

The logistic regression classifier performed the best across all the different cuts and labellings of the data, in particular for D2 & D4. However, this is due to the fact that there is a low frequency of other 'Non-AD' or 'Control' labels, hence the relative low recall and F1 scores. Instead of removing the less common target entries, it was decided to base the analysis for this project on D4, where all 'Non-AD' or 'Control' targets were removed and both severities ('Possible' and 'Probable') of AD were merged to return a target label of 'All AD'.

Interactional Features Selection

A total of 24 interactional features were encoded originally. The feature selection process for these features was done by ranking all 24 encoded features based on their ANOVA F-value between label/feature for the classification task of predicting the diagnosis. Using Python's SelectKBest (sklearn) function where the number of features that are being ranked, k, is inputted as a parameter. In this case, k is equal to 26. It was decided that the top interactional features (Top IF's) would be selected based on their ANOVA F-value score (predicting weight) which is plotted in Appendix Figure 3. All features with a non-zero ANOVA F-value score were selected as the Top IF's therefore the only feature that was removed was 'Interruption_Q' which was encoded to represent if the participant interrupted either themselves or the examiner in the form of a question. Therefore, Top IF's were made up of 23 features (see Appendix Table 2).

From Appendix Figure 3, it can be seen that the top two predicting features based on ANOVA F-Value all appeared to fall under the 'Conversational' umbrella feature that was created. The ratio of utterance starts between the participant and examiner (Turn-Taking Ratio) and a normalised count of questions asked by the invigilator, which can both be said to represent the examiner needing clarify with the participant that they comprehend the task at hand or suggests that the examiner was required to encourage them for more information. The fact that the third highest predicting variable (Incomplete Word Count of the participant) suggests that this could be the cause for the need for the examiner to have to ask more questions, and so have a higher utterance start rate. The combination of these three features could represent the amount of involvement that the examiner needed to manifest throughout the task.

Further exploration of the interactional features was carried out by investigating the correlation between the target variable and the each of the Top IF's.

Exploration of Correlation between Top Interactional Features and Diagnosis

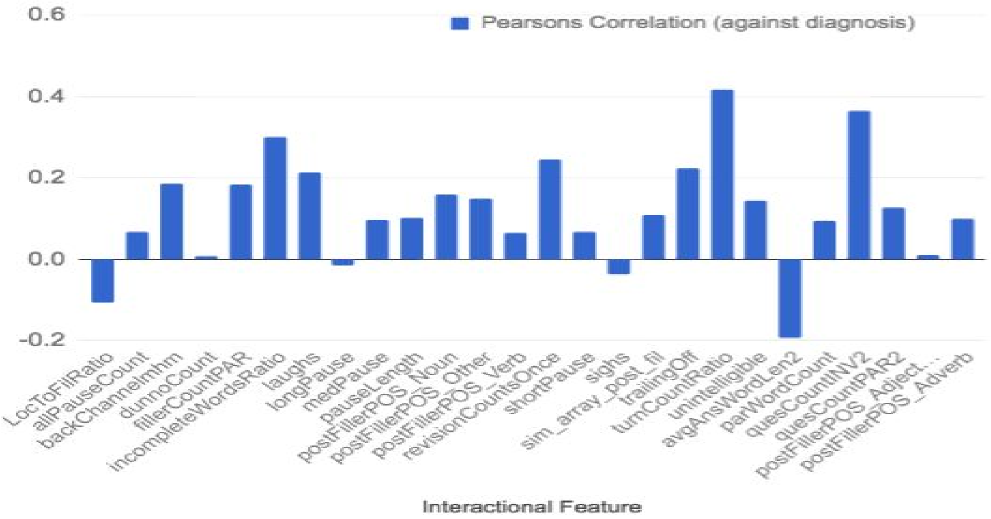


Fig 1. Correlation between target variables ('Probable AD') and the top interactional features.

The aim of this section[4] was to investigate how each of the interactional variables correlate with the target variables (diagnosis). The also acts as a sense check that the variables were in encoded as expected. The strength and direction (positive or negative) of each variable's correlation with the target was investigated. To ensure a representative output, it was decided to compare the control group against the group with the highest level of dementia (Probable AD) using Pearson's correlation where if a variable has a positive correlation, it's value increases in the direction of 'Probable AD' and if it has a negative correlation, it's value increases in the direction of the control target.

[4] All data in the section refers to Fig. 3 (above) unless stated otherwise.

The two variables that have the strongest positive correlation (TurnCountRatio, Examiner's Question Frequency) both suggest that the examiner needed to manifest a stronger presence throughout the session, encouraging the participant for more information or having to clarify that the participant understands. This theory is strengthened by the fact that 'backChannelhmh' which represents is the normalised count of backchannels used by the examiner also positively correlates with the diagnosis, and overlaps with the reasoning behind the variable with strongest negative correlation (average answer word length). It suggests that the longer the length of the participants sentence, the more likely they are to not have 'Probable AD'.

The number of fillers (normalised by being divided by the total words) uttered by the participant positively correlates with the participant having AD. This is also said for the 'revision count' variable, which can be said to roughly represent self-repair. Since it positively correlates with the diagnosis, it suggests that the AD patients recognise when they say the unintended word that needs to be repaired and attempt to repair it.

The slightly negative correlation of the 'locToFillerRatio' variable and the target suggests that there is some relationship between the two although not as strong as the previous variables mentioned. It suggests that the earlier on in a sentence (lower locToFillerRatio value) that the filler occurs, the more likely the participant is to have AD.

The 'Probable AD' diagnosis correlates with participants uttering a filler, which can be used as a sign of hesitation, before nouns and adverbs, where as there is little correlation with the target variables and a patient utters a filler before an adjective, which was expected as it has be shown in [26] that AD patients utter few nouns that other POS tags however, it has now been shown that this is also through for the POS tags that occur after a filler.

Evaluation Methods

To evaluate the performance of interactional features being used to computationally diagnose Alzheimer's types and other dementia, the interactional features would be merged with variables that were already known to be strong predictors of the diagnosis (ie. the top non-IF's), to see if any of the interactional features had a higher predicting weight (ANOVA F-Value) than any of the non-IF's. If any of the interactional features outperformed the non-interactional features regarding their predicting weight score, it can be said that the combination of both interactional and non-interactional features predict better than that of just non-interactional features.

The null hypothesis was that the top predicting variables regarding the computational diagnosis of Alzheimer types are based on non-interactional, linguistic based phenomenon. These are represented by the top 23 non-interactional features (Top Non-IF's) that were created based on the non-interactional features in Fraser et al. [2]. The alternative hypothesis is that the addition of other variables, encoded based on the interactions of the patient, are better predictors and can be used to improve the classification of diagnosis. This is the outcome if any of theTop IF's fall into the top 23 variables based on predictive weight when both the Top Non-IF's and Top IF's predictive weights are compared. The 'new' top 23 variables based on predictive power will then be used to train the same classifier and same data that was built earlier using the top Non-IF's. If there is an improvement in performance scores, it cements the statement that the combination of both interactional and non-interactional features predict better than that of just non-interactional features regarding the computational diagnosis of Alzheimer’s disease.

Results

The final top 23 predicting features were made up of 52.2% interactional and 47.8% non-interactional. This combination of features improved the originally all non-interactional classifier accuracies by 4.96%, with an accuracy score of 82.46%, recall at 78.16% and an F1 Score of 79.36% (See Appendix Table 3).

In terms of the interactional variables ‘Umbrella Features’ and the non-interactional variables ‘Factor Groupings’, the final top 23 variables consisted of 17.4%: ‘Fillers’, 30.4%: ‘Semantic Impairment’, 17.4%: ‘Information Impairment’, 13.0%: ‘Conversation’ and 17.4%: ‘Unintentional Silence’. (See appendix Fig 6)

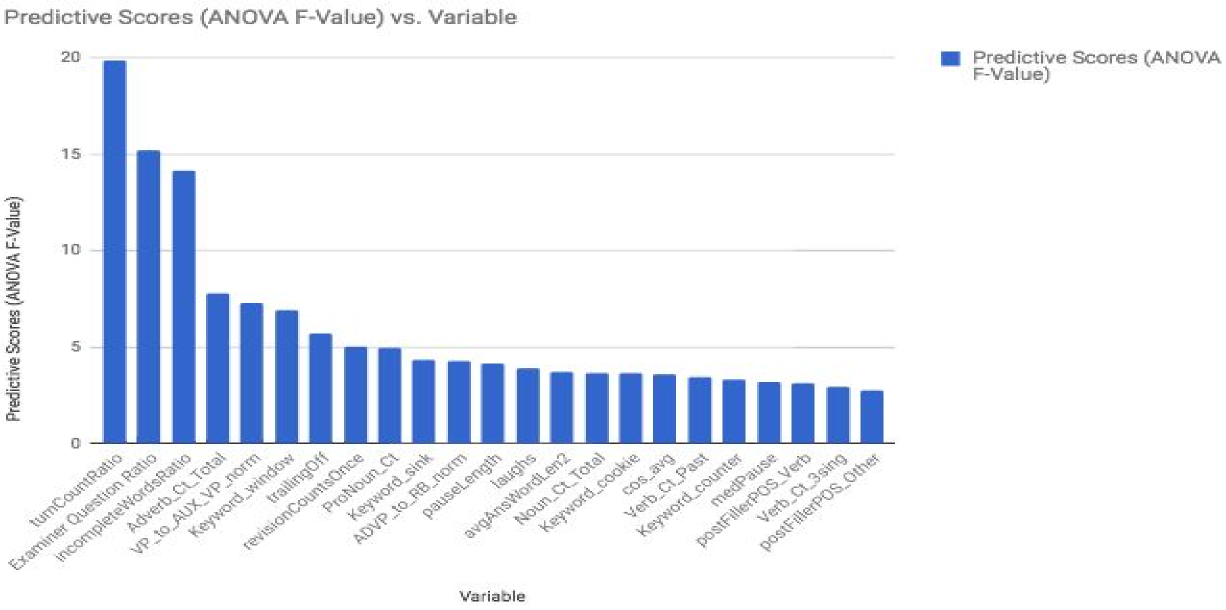


Fig 2. Predictive weights of ‘New Top 23’ Features (data from Appendix Table 6)

There are three variables that have a significantly higher predicting scores (ANOVA F-Value) than the remaining features which are Turn-Take Ratio (TTR), Examiner Question Ratio (EQC) and Incomplete Word Ratio (ICR). The combination of these three variables having a significantly higher predictive scores suggest that the phenomenon of the examiner being required to manifest a role of higher dialogue and presence can be represented by encoding these features and used to aid the computational diagnosis of AD.

Discussion

It is clear from the above results that the combination of features that represent both linguistic and interactional phenomenon perform better at computationally diagnosing AD than using features that are solely based on linguistics. This is due to the fact that, as hypothesized earlier on the in project, by adding in an extra layer of known symptoms to the classifier, it will improve predictions and classification of AD.

Regarding the target (diagnosis) predictive weight score the two highest ranking variables, TTR (defined as ‘examiner utterance starts’/’participant utterance starts’) and a count of the questions the examiner stated (EQR), were both calculated based on the amount of involvement and presence the examiner was required to manifest throughout the session, with predictive scores of 19.85 and 15.18 respectively. The third highest predicting variable, with a predictive score of 14.14, represents a count of incomplete words (IWR) uttered by the participant throughout the session. When compared with the correlation analysis (interactional features against the diagnosis variable) carried out previously, it can be seen that these three top predicting features all positively correlate with the diagnosis of ‘Probable AD’. A suggested reasoning behind this is because the examiner has a need to take on a role with a higher level of presence (higher TTR) throughout the session due to the fact that the participant is unclear (higher IWR) leading to the examiner requiring to ask questions (higher EQR & TTR) in order to clarify what the participant is intending to express and to encourage the participant to keep going. The combination of these three variables having a significantly higher predictive score may be used to suggests that the phenomenon of the examiner being required to manifest a role of higher dialogue and presence can be represented by encoding these features and similar features to these, in order to aid the computational diagnosis of AD. This leads to the hypothesis that, if encoded correctly, conversation and dialogue analysis may play a significant role in computationally classifying Alzheimer’s type. This could be proven by investigating dialogue based features.

Conclusion

Since interactional features replaced over half of the originally all non-interactional top predicting features regarding diagnosis predicting weight, as well as causing an improvement in the classifier's accuracy, one can conclude that encoding interactional features, in particular dialogue based features that represent the amount of involvement the invigilator is required to manifest throughout the session, in addition to non-interactional features, can assist in computationally classifying Alzheimer's disease.

Further Work

The results suggest that dialogue plays a key role in the classification of AD therefore further investigation would include exploration of variables created based on the dialogue between the examiner and the participant. This would include running the data through a dialogue act tagger and carrying out an analysis similar to the above, including a correlation analysis with the target variable.

It would have also been beneficial to investigate a different data set, other than recording of the “Boston Cookie Theft” task as other psychological tasks may include a stronger level of interactions between the examiner and the patients. It was found that there are a number of other tests available (Sentence, Recall, Fluency) on Dementia Bank however, there is no control group. Therefore, in order to ensure a fair analysis of these data sets, a proxy control group would have to be created. It would be interesting to carry out sentiment analysis on a dataset that expresses patients emotion. This alone could be used as a diagnosis classifier variable. It would also be interesting to investigate if the main sentiment of the patient changes throughout the session and well as investigating when it changes and in relation to what – ie. What interaction did the participant have with the examiner to cause this change in sentiment?

Self-repair is a fascinating aspect of the human language and plays a significant role when in the process of word-searching (anomia). This is a key symptom of dementia due to aphasia. The self-correction variable in this project appeared in the top ten predicting features. It was also felt however, that the definition of the self-correction (based on CHAT annotation of ‘self restarts’) variable could be made more robust by running the data through a high end self-repair detector (STIR). It is felt that a number of different variables could be built on repair including finding the cosine difference between word that needed repairing and intended words. This could be said to measure the mistake made.

This project focuses on diagnosis Alzheimer’s disease, and even though this is the most common form of dementia, it would be interesting to investigate what features predict other NON-AD dementia (MCI, Vascular, Memory). As was seen earlier on in the project, this dataset is very unbalanced regarding non-AD target labels. Therefore, one could use an anomaly detection approach as opposed to classification and investigate the features that aid the prediction.

Bibliography

1: Becker, J. T., Boller, F., Lopez, O. L., Saxton, J., & McGonigle, K. L., The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis., 1994

2: Fraser, K.C., Meltzer, J.A. and Rudzicz, F., Linguistic features identify Alzheimer’s disease in narrative speech., 2016

3: Alzheimer's Society Website, Alzheimer's Society - What is Dementia?, 2017

4: Svennevig, J. and Lind, M., Dementia, interaction, and bilingualism: An exploratory case study., 2016

5: Alzheimer's Society Website, Facts for the media - Alzheimer's Society., 2017

6: Croisile, B., Ska, B., Brabant, M.J., Duchene, A., Lepage, Y., Aimard, G. and Trillet, M., Comparative study of oral and written picture description in patients with Alzheimer's disease., 1996

7: Guinn, C.I. and Habash, A., Language Analysis of Speakers with Dementia of the Alzheimer's Type., 2012

8: Bucks, R.S., Singh, S., Cuerden, J.M. and Wilcock, G.K., Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance., 2000

9: Alzheimer's Society Website, Alzheimer's Society - Behaviour Changes, 2017

10: Carozza, L.S. ed., Communication and Aging: Creative Approaches to Improving the Quality of Life., 2015

11: Giles, E., Patterson, K. and Hodges, J.R., Performance on the Boston Cookie Theft picture description task in patients with early dementia of the Alzheimer's type: missing information., 1996

12: Bird, H., Ralph, M.A.L., Patterson, K. and Hodges, J.R., The rise and fall of frequency and imageability: Noun and verb production in semantic dementia., 2000

13: Silvast, M., Aphasia therapy dialogues., 1991

14: Fergadiotis, G. and Wright, H.H., Lexical diversity for adults with and without aphasia across discourse elicitation tasks., 2011

15: Szatloczki, G., Hoffmann, I., Vincze, V., Kalman, J. and Pakaski, M., Speaking in Alzheimer's disease, is that an early sign? Importance of changes in language abilities in Alzheimer's disease., 2015

16: MacWhinney, B., The TalkBank Project., 2007

17: Becker JT, Boiler F, Lopez OL, Saxton J, McGonigle KL, The natural history of Alzheimer's disease: Description of study cohort and accuracy of diagnosis., 1994

18: Goodglass H, Kaplan E, The Boston Diagnostic Aphasia Examination., 1983

19: Mayo Foundation for Medical Education and Research (MFMER), Mild Cognitive Impairment (MCI), 2016

20: Alzheimer's Society Website, Vascular Dementia, 2017

21: MacWhinney, B., The CHILDES Project: Tools for analyzing talk, 3rd edition., 2000

22: Bloom, L., Lifter, K. & Hafitz, J., Semantics of verbs and the development of verb inflection in child language., 1980

23: Itakura, H., Describing conversational dominance., 2001

24: Lesta, B. and Petocz, P., Familiar group singing: Addressing mood and social behaviour of residents with dementia displaying sundowning., 2006

25: Cummings, J.L., Benson, D.F., Hill, M.A. and Read, S., Aphasia in dementia of the Alzheimer type., 1985

26: Jarrold, W., Peintner, B., Wilkins, D., Vergryi, D., Richey, C., Gorno-Tempini, M.L. and Ogar, J., Aided diagnosis of dementia type through computer-based analysis of spontaneous speech., 2014

Appendix

Fig 1: Correlation Matrix of All NI Features (Pre-FS)

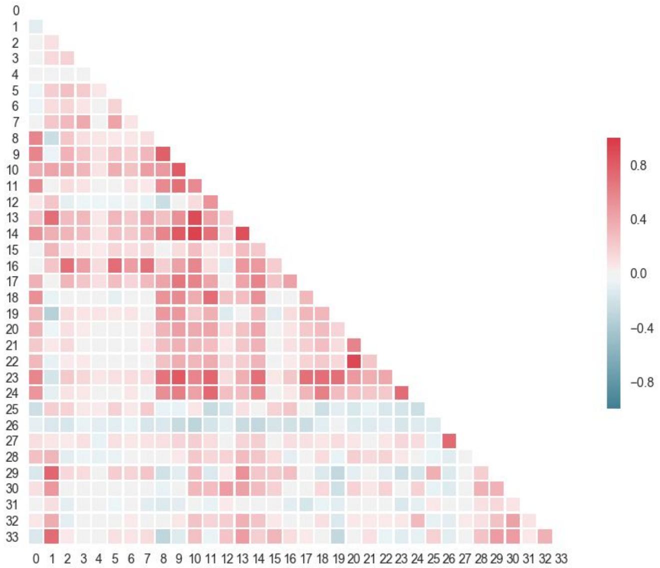


Fig 2: Correlation Matrix of Remaining NI Feature’s (Post-FS)

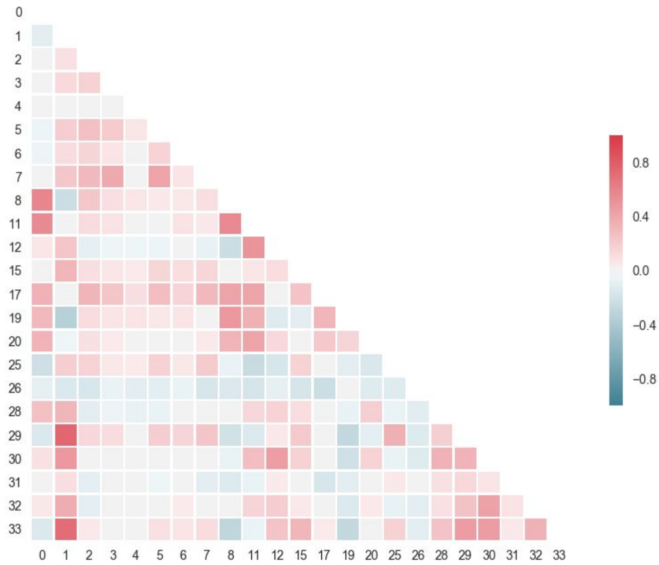


Table 1. Fig. 1 and Fig. 2 legend & list of NI variables encoded

0: 'Adverb_Ct_Total'	12: 'ProNoun_Noun_Ratio'	24: 'Verb_Ct_non3sing'
1: 'Avg_Word_Len'	13: 'Script_Letter_Ct'	25: 'cos_avg'
2: 'Keyword_cookie'	14: 'Script_Word_Ct'	26: 'wordsOnceOverTotal'
3: 'Keyword_counter'	15: 'Subj_IU_Total'	27: 'wordsUsedOnce'
4: 'Keyword_curtain'	16: 'Total_Keywords_Mentioned'	28: ‘ADVP_to_RB_norm’
5: 'Keyword_sink'	17: 'Verb_Ct_3sing'	29: ‘NP_to_DT_NN_norm’
6: 'Keyword_stool'	18: 'Verb_Ct_Base'	30: ‘NP_to_PRP_norm’
7: 'Keyword_window'	19: 'Verb_Ct_Gerund'	31: ‘ROOT_to_FRAG_norm’
8: 'Noun_Ct_Total'	20: 'Verb_Ct_Past'	32: ‘VP_VBG_norm’
9: 'PAR_UTT_ct'	21: 'Verb_Ct_Past_P'	33: ‘VP_to_AUX_VP_norm’
10: 'PAR_Word_Ct'	22: 'Verb_Ct_Past_T'	
11: 'ProNoun_Ct'	23: 'Verb_Ct_Total'	

**PAR: Participant, Ct: Count, UTT: Utterance, IU: Information Unit*

Table 2. List of all interactional features encoded.

0: Interruption_Q	8: longPause	16: sighs
1: LocToFilRatio	9: medPause	17: sim_array_post_fil
2: allPauseCount	10: pauseLength	18: trailingOff
3: backChannelmhm	11: postFillerPOS_Noun	19: turnCountRatio
4: dunnoCount	12: postFillerPOS_Other	20: unintelligible
5: fillerCountPAR	13: postFillerPOS_Verb	21: avgAnsWordLen2
6: incompleteWordsRatio	14: revisionCountsOnce	22: quesCountINV2
7: laughs	15: shortPause	23: quesCountPAR2

'0: Interruption_Q' was the only feature removed during the feature selection process.

Fig. 3 Predictive Scores (X-axis) vs. All IF's (Y- axis)

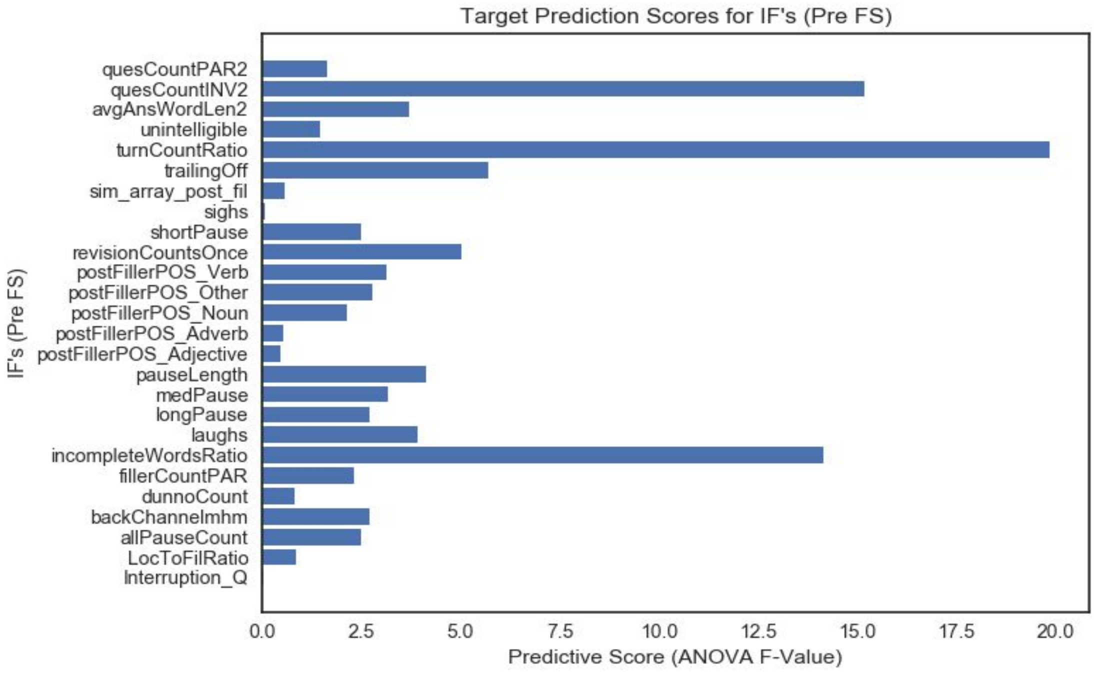


Table 3. Performance scores of all classifiers on different cuts of the data.

X = Top NIF's (Count 23)		'NIF Only' Performance (10-fold CV)			'NIF & IF' Performance (10-fold CV)			Absolute difference (from 'NIF only' to 'NIF & IF')		
y	Classifier	Precision	Recall	F1 Score	Precision	Recall	F1 Score	Precision	Recall	F1 Score
D1:	DT	0.5233	0.4727	0.4917	0.6275	0.5364	0.5645	10.42%	6.36%	7.27%
	LR	0.5881	0.6200	0.5948	0.6367	0.6818	0.6506	4.87%	6.18%	5.58%
	GNB	0.5999	0.1236	0.1682	0.6568	0.3436	0.4298	5.69%	22.00%	26.17%
	KNN	0.5277	0.5145	0.5046	0.5620	0.5145	0.5110	3.43%	0.00%	0.64%
D2:	DT	0.6197	0.5670	0.5810	0.6680	0.5909	0.6157	4.83%	2.39%	3.47%
	LR	0.6991	0.6831	0.6821	0.7626	0.7573	0.7523	6.35%	7.42%	7.02%
	GNB	0.7587	0.2306	0.3096	0.7454	0.4627	0.5377	-1.33%	23.21%	22.81%
	KNN	0.6136	0.5789	0.5838	0.6475	0.5770	0.5904	3.40%	-0.19%	0.66%
D3:	DT	0.5160	0.4636	0.4816	0.6014	0.5200	0.5432	8.54%	5.64%	6.15%
	LR	0.5889	0.6200	0.5953	0.6344	0.6818	0.6496	4.55%	6.18%	5.42%
	GNB	0.5983	0.1309	0.1745	0.6538	0.3873	0.4605	5.55%	25.64%	28.60%
	KNN	0.5346	0.5109	0.5044	0.5571	0.5145	0.5124	2.25%	0.36%	0.80%
D4:	DT	0.7117	0.6293	0.6517	0.7414	0.6836	0.7006	2.97%	5.43%	4.90%
	LR	0.7750	0.7233	0.7381	0.8246	0.7816	0.7936	**4.96%**	**5.82%**	**5.54%**
	GNB	0.7488	0.6914	0.7094	0.8023	0.7333	0.7496	5.35%	4.19%	4.03%
	KNN	0.7102	0.6231	0.6455	0.6916	0.5791	0.6043	-1.86%	-4.39%	-4.12%

Table 4. Count of data for the different labellings of the target variable for Appendix Table 3 (above).

	Control	Possible AD	Probable AD	MCI	Vascular	Memory
D1:	241	21	237	43	5	3
D2:	241	21	237	0	0	0
D3:	241	21	237	51 ('Other')		
D4:	241	258 ('All AD')		0	0	0

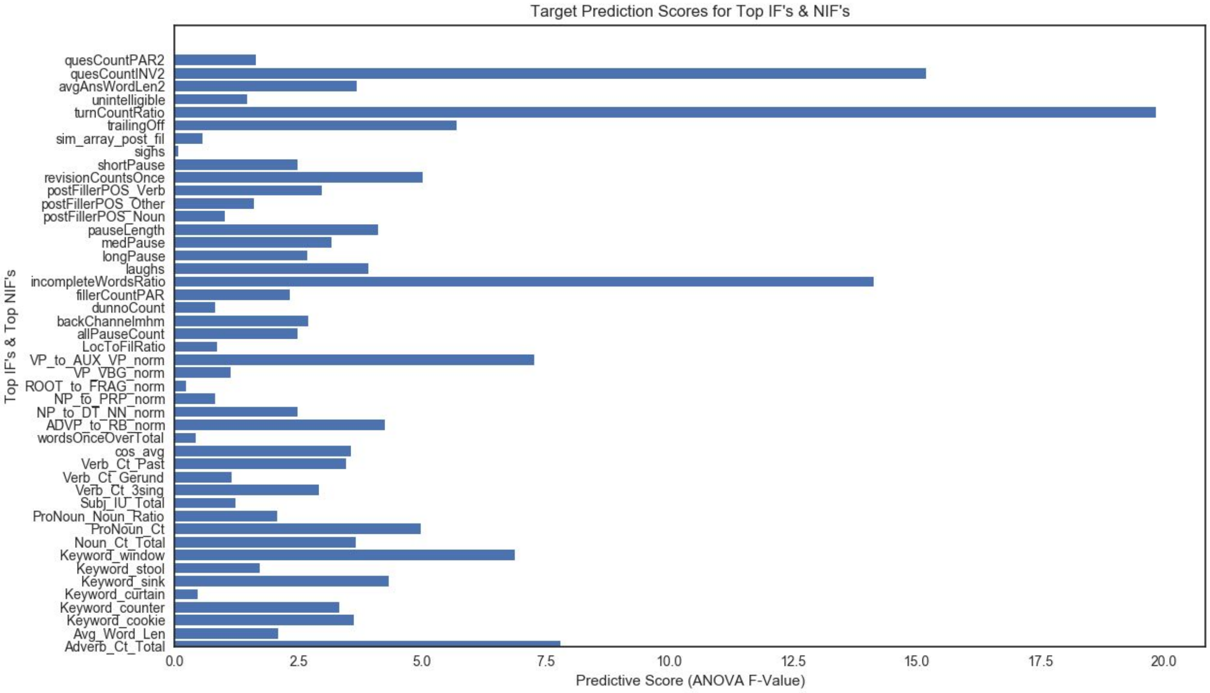


Fig. 4 Predictive Scores (X-axis) vs. Top IF's and NIF's (Y- axis)

Table 5. POS Tags - High Level Groupings

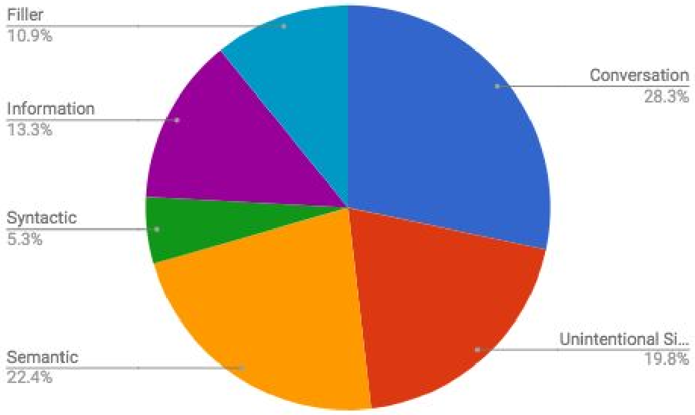
POS Tag	Description	Grouping
CC	Coordinating conjunction	Other
CD	Cardinal number	Other
DT	Determiner	Other
EX	Existential there	Other
FW	Foreign word	Other
IN	Preposition or subordinating conjunction	Other
JJ	Adjective	Adjective
JJR	Adjective, comparative	Adjective
JJS	Adjective, superlative	Adjective
LS	List item marker	Other
MD	Modal	Other
NN	Noun, singular or mass	Noun
NNS	Noun, plural	Noun
NNP	Proper noun, singular	Noun
NNPS	Proper noun, plural	Noun
PDT	Predeterminer	Other
POS	Possessive ending	Other
PRP	Personal pronoun	Other
PRP$	Possessive pronoun	Other
RB	Adverb	Adverb
RBR	Adverb, comparative	Adverb
RBS	Adverb, superlative	Adverb
RP	Particle	Other
SYM	Symbol	Other
TO	*to*	Other
UH	Interjection	Other
VB	Verb, base form	Verb

VBD	Verb, past tense	Verb
VBG	Verb, gerund or present participle	Verb
VBN	Verb, past participle	Verb
VBP	Verb, non-3rd person singular present	Verb
VBZ	Verb, 3rd person singular present	Verb
WDT	Wh-determiner	Other
WP	Wh-pronoun	Other
WP$	Possessive wh-pronoun	Other
WRB	Wh-adverb	Other

Table 6. Final Top 23 Features and their Predictive Weights

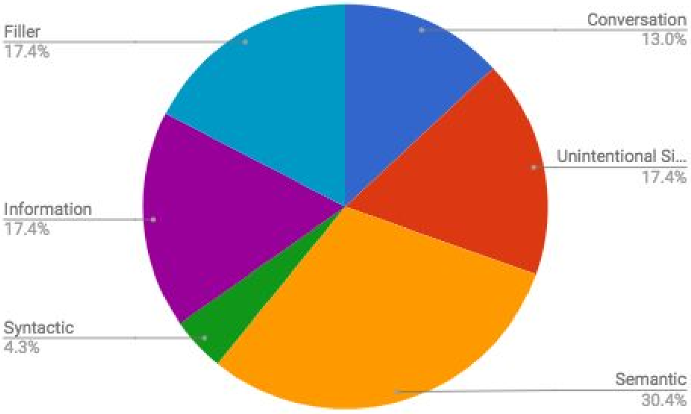
Predictive Scores (ANOVA F-Value)	Variable	Feature Type	Factor/Umbrella
19.84673728	turnCountRatio	IF	Conversation
15.1849209	Examiner Question Ratio	IF	Conversation
14.13549536	incompleteWordsRatio	IF	Unintentional Silence
7.796872585	Adverb_Ct_Total	NIF	Semantic
7.275023488	VP_to_AUX_VP_norm	NIF	Syntactic
6.873187562	Keyword_window	NIF	Information
5.700162083	trailingOff	NIF	Unintentional Silence
5.015922563	Self-Repair (revision count)	IF	Filler
4.969879595	ProNoun_Ct	NIF	Semantic
4.323192073	Keyword_sink	NIF	Information
4.245024757	ADVP_to_RB_norm	NIF	Semantic
4.114834213	pauseLength	IF	Unintentional Silence
3.912345997	laughs	IF	Filler
3.687530286	avgAnsWordLen2	IF	Conversation
3.653043837	Noun_Ct_Total	NIF	Semantic
3.629273321	Keyword_cookie	NIF	Information
3.562162496	cos_avg	NIF	Semantic
3.458298487	Verb_Ct_Past	NIF	Semantic
3.33526023	Keyword_counter	NIF	Information
3.162666096	medPause	IF	Unintentional Silence
3.145025579	postFillerPOS_Verb	IF	Filler
2.922614281	Verb_Ct_3sing	NIF	Semantic
2.770281656	postFillerPOS_Other	IF	Filler

Fig 5. Chart based on Table 6 data (Predictive Weight grouped by Factor/Umbrella)



\

Fig 6. Chart based on Table 6 data (Count of final features grouped by Factor/Umbrella) ****Ignoring weight value***


Filler
Conversation
Unintentional Si...
Information
Syntactic
Semantic