**Speech Deterioration in Alzheimer’s disease**

**Some signs of the disease which can be analysed through speech analysis:-**

* confusion;
* memory loss
* getting lost in familiar areas;
* personality changes;
* depression (as the person recognizes his or her deficits);
* significant memory loss;
* difficulty following simple directions;
* decreasing communication skills;
* By the final stages speak intelligibly.
* Decrease in vocabulary
* Reduction in word specificity
* Increase in indefinite nouns and low-imageability verbs
* Noun to verb ratio changes as more low-image verbs are used.
* Use of more agentless passive (eg- John was fired)

1. **Learning predictive linguistic Features for Alzheimer’s Disease and related Dementia’s using verbal Utterances**

Using Machine Learning algorithms to develop models using the syntactic and lexical features of the verbal utterances.

Source of data was the Dementia Bank data.

Parsing of the Dementia Bank dataset using the Stanford Parser and then creating a syntactic tree structure of each sentence.

Features used:-

1. Coordinated sentences: these are sentences whose clauses are connected using coordinating conjunctions. POS tagger:- CC
2. Subordinated sentences: these are sentences that are subordinate to the independent primary sentences to which they are linked. POS tagger:- S
3. Reduced sentences: those subordinate sentences without a conjunction but with nominal verb forms. POS tagger: - VBG and VBN.
4. Number of predicates: the number of predicates can be determined to measure sentence complexity. The predicates were extracted using and algorithm which locates transitive verbs which are followed by one or more arguments.
5. Average number of predicates
6. Dependency distance: to measure the grammatical complexity. It is equal to summation of all dependency distances where dependency distance=absolute difference between the serial position of two words that participate in a dependency relation.
7. Number of dependencies: The number of unique syntactic dependencies in a sentence were examined.
8. Average dependencies per sentence: - Average number of dependencies per sentence.

**NB- Dependency is the notion that linguistic units, e.g. words, are connected to each other by directed links. The (finite) verb is taken to be the structural center of clause structure. All other syntactic units (words) are either directly or indirectly connected to the verb in terms of the directed links, which are called *dependencies***.

1. Production Rules: Production rules derived from parse trees has been explored in a number of NLP classification texts. We count the number of unique production rules in the CFG extracted from each patients’ narrative.

Lexical features (revision and repetition features): -

1. Utterances: To count the number of utterances per patient. An utterance is the beginning of a verbal communication to the verbal pause. There can be many utterances in a sentence. An utterance forms a grammatical lexicon for a language. Thus it shows the language strength of a potential patient.
2. Mean length of Utterances (MLU): ratio of the total number of words to the number of utterances.
3. Function words: To count the total number of function words.
4. Unique Words: To count the total number of unique words.
5. Word Count: To count the total number of words.
6. Character Length: The absolute character length of the patient’s narrative.
7. Total sentences: To count the total number of sentences.
8. Repetitions: Number of immediate word repetition in the patient’s narrative.
9. Revisions: Count of the pause positions where the patient retraced the preceding error and made corrections.
10. Lexical Bigrams: Number of unique bigrams**. (NB:- A bigram or digram is a sequence of two adjacent elements from a** [**string**](https://en.wikipedia.org/wiki/String_%28computer_science%29) **of** [**tokens**](https://en.wikipedia.org/wiki/Token_%28parser%29)**, which are typically letters, syllables, or words. A bigram is an** [***n*-gram**](https://en.wikipedia.org/wiki/N-gram) **for *n*=2. The frequency distribution of every bigram in a string is commonly used for simple statistical analysis of text in many applications, including in computational linguistics, cryptography, speech recognition, and so on.)**
11. Morphemes: To measure the number of morphemes. A morpheme is a word or a part of it that can’t be further subdivided.(In [linguistics](https://en.wikipedia.org/wiki/Linguistics), a **morpheme** is the smallest grammatical unit in a language. In other words, it is the smallest meaningful unit of a language.)

All the Machine Learning Algorithms were implemented using the Weka toolkit with the default settings.

Conclusions:-

1. Disease group has more difficulty in constructing complex sentences unlike the healthy group.
2. Disease group performed more immediate word repetitions and made more revisions on grammatical errors in their narrative.
3. Disease group tend to do more utterances.

Future Work suggested:-

Training the data on a larger dataset which could lead to more accurate results.

1. **Linguistic Features Identify Alzheimer’s Disease in narrative speech**

1. Analysis of narrative speech sample using machine learning algorithms.

2. Heterogeneity of linguistic impairment in the participants of AD.

Data derived from the Dementia Bank corpus.

Only the word level transcription was kept along with the utterance segmentation

Features:-

* 1. Parts of Speech: Used the Stanford tagger. Computing the frequencies of the different parts of speeches, normalized by the total number of words in each utterances. Ratios were also computed e.g. - nouns to verbs, pronouns to nouns etc.
  2. Syntactic Complexity- A parse tree was formed using the Stanford parser. Then syntactic complexity was measured by using the mean length of sentences, T units and clauses and calculations from the parse tree including the height of the tree and the mean, depth of the tree (embeddedness).
  3. Grammatical constituents: CFG features. CFG can help to mark possible syntactic differences.
  4. Psycholinguistics: Frequency of highly familiar words.
  5. Vocabulary richness: TTR, Brunet’s index, Honore’s statistic.
  6. Information Content:
  7. Repetitiveness:Repetitive content by measuring the average distance between repetitions.
  8. Acoustics
  9. Identification of AD by using Machine learning
  10. Factor Analysis: Semantic impairment, acoustic abnormality, Syntactic impairment, information impairment

1. **Automatic detection and Rating of Dementia of Alzheimer type through Lexical analysis of Spontaneous speech**

Computational approach to measure the level of impairment in patients.

Symptoms of the disease:

* + 1. Breakdown in semantic processing
    2. Shallow vocab
    3. Word finding difficulties leading to deterioration of speech

Ways to identify the level of impairment on patients:-

1. The Mini Mental State Exam- The patients will answer some question to test the mental condition of the patients.
2. Verbal picture descriptions- To verbally describe the line diagrams kept in front of them.

The Lexical Approach:-

Some features of the spontaneous words of the patients were considered:

1. Noun(N) rates
2. Pronoun(P) ‘’
3. Adjective(A) ‘’
4. Verb(v) ‘’
5. Type token ration(TTR) = ratio of the total vocabulary V to the total text length N
6. Brunet’s Index(W) = N^(V^(-0.165))
7. Honore’s Statistic(R) = (100 log N)/ (1-(V1/V)) where V1 = the number of words spoken only once.
8. Clause-like semantic unit (CSU) = ability to form noun and verb phrases and gives an indication of the flow of speech. Hand tagged data was used to count cohesion boundaries in phrases. Patients would not be able to form long phrases thus increasing the CSU rate.
9. Common n-grams approach (CNG) : This approach uses character n-grams to model consistencies in author style. This approach overcomes the problems of word based model such as word dependencies, and word sparsity due to large vocabulary.
10. Classification via Common word frequencies: Most common words were treated as style markers.

This research used the ACADIE dataset (study of donepezil). The research concluded that several ML algorithm and natural language processing tools can help us identify the level and degree of impairment in Alzheimer’s patients.

**4. Aided Diagnosis of Dementia type through Computer Based Analysis of Spontaneous Speech**

To identify the dementia type by using the speech recordings of patients.

Features used for Lexical feature extraction (LFE):- Frequencies of nouns, verbs, function words, words about emotion, etc.

Two types of computer based lexical analysis:-

1. Frequencies of the different Parts of Speech using the POS tagger.
2. Dr. Pennebaker’s Linguistic inquiry and word count (LIWC) software which determines words in different categories (emotional or cognitive etc.)

The LIWC is an extremely strong tool to analyze text data but it is not free.  
Three machine learning methods were used:-

1. Logistic regression- a statistical learning technique for determining categorical outcomes.
2. Multi layered Perceptrons – an AI learning method that roughly mimics biological neural networks.
3. Decision trees – another AI technique which induces sets of rules to predict the outcome.

Tool used:- Weka.

5. **Computational cognitive modeling of inflectional verb morphology in Spanish speakers for the characterization and diagnosis of Alzheimer’s disease**

Method used:-

1. Finding the task that exhibits behavioral differences between healthy and impaired subjects
2. Preparing the computational cognitive architecture with the knowledge to deal with the selected task
3. Modelling the individual subject’s behavior to obtain the parameters of the architecture specific to each participant
4. Applying machine learning techniques on the information given by the cognitive models to learn the classification model that supports impairment diagnosis.

Attributes taken into account:-

1. RT- Retrieval threshold
2. ANS- noise introduced into the memory retrieval process
3. BLL- forgetting factor

6. **Tracking Discourse Complexity Preceding Alzheimer’s disease Diagnosis: A case Study comparing the press conferences of Presidents Ronald Reagan and George Herbert Bush**

Data used:-

Presidential transcripts collected from the American Presidency Project (APP) archive as a data source for this project-

1. A prepared statement read by the president
2. A spontaneous question answer session in which the press asked a number of question to which the president answered.

Features used:-

1. Number of unique words
2. Number of fillers
3. Number of Non-specific nouns
4. Number of low-imageability verbs (Imageability is characterized by the ability of a word to generate images in the mind).

Tool used:-

The Natural language Processing Toolkit (NLTK) in python.

**7. Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three British novelists**

The works of three British novelists were analyzed namely:-

Iris Murdoch, Agatha Christie, and PD James.

Data used: - Twenty of the novels of each of these novelists were collected and examined.

1. Lexical Analysis: - Simple linear regression of the features was considered against the authors’ age and a statistically significant relationship was built between the age and the value of the measure.

Features used (vocabulary size):-

1. Vocabulary size using the TTR (type token Ratio) = ratio of vocab used to the total text length.
2. Word type introduction rate (WTIR) = cumulative number of unique lemmatized types computed at every 10,000-token interval.
3. Lexical repetition:-
4. Counting global word n-gram repetitions.
5. Counting local repetitions.
6. Lexical specificity- computing proportions of indefinite nouns, LI verbs tokens in each texts. A higher proportions indicate dependency on generic words.
7. Word class deficit- proportions of each word class over the entire length of each text, in terms of both word tokens in order to look for signs of deficit to measure vocab size. Measurement of the changes in the pattern of nouns, pronouns, verbs etc.
8. Noun tokens
9. Verb tokens
10. Percent of verb types
11. Percent of pronouns
12. Percent of noun tokens
13. Percent of common noun types
14. Percent of adverb type and tokens.
15. Common noun
16. Proper noun
17. Noun
18. Content verb
19. Adjective
20. Adverb
21. Pronoun
22. Fillers
23. Syntactic analysis
24. Mean length of utterance (MLU) and mean number of clauses per utterance (MCU) was calculated for each sentence parse tree.
25. Parse Tree depth and Yngve depths- to measure the degree of embeddedness of the sentences so as to get an idea of the syntactic complexity.
26. D-level
27. Passive Voice- Frequency of passive voice structures.

8. **The effects of very early Alzheimer’s disease on the characteristics of writing by a renowned author**

Three of the books of Iris Murdoch were analyzed. Those three books were written at three different phases of her career. “The Jackson’s dilemma being her last book”.

Features:-

1. Percentage of the Parts of speeches
2. Values of lexical variables
3. Syntactic complexity (mean number of words or clauses)
4. Vocabulary (repetition of words)

**My proposed Idea**

So as we have seen from our literary articles the features which have been used all these years for Alzheimer’s analysis using the textual transcripts include-

1. Frequency of the Parts of speeches
2. Syntactic complexity
3. Vocabulary

Most of the papers have focused on these features with some modifications in individual papers.

Steps we can follow:-

1. Tokenization of the texts.(using NLTK)
2. Stemming of the texts to one canonical form. (Porter’s algorithm can be used).
3. Then the following steps can be repeated for each document:-
4. Calculating the frequency of the Parts of speeches( Stanford Tagger)
5. The weight of each POS can be calculated using term Frequency with Inverse Document Frequency (TF-IDF)

W=tf\*log (n/df)

tf=Number of occurrences of a particular POS in a document

n=Total number of documents considered

df=term frequency of that word in the range of documents considered.

1. Each POS can be treated as particles of the Swarm. We can cluster documents using the particle swarm optimization technique.Initially each particle will be treated as the centroid of the cluster. The particles will move towards a solution based on its experience and the experience of its neighbors. This movement will depend on the value of the fitness function (which we will have to generate). Finally the swarm will converge at a solution.
2. The solutions from the different documents can be evaluated for a particular patient over the years which can help us identify the trends in the deterioration. The trends can be changes in the patterns of usage of different Parts of Speeches. This idea can be extended to judge syntactic complexity or vocabulary.

**Steps:-**

# Make text files clean without any of ~@#$%^&\*\_-+=[{]}|\<>/`

# Store the word lists and sentence lists separately for each file in each folder

f

# 1. Lowercase transformation, Stemming, Unique word sets construction

# 2. Lexical Analysis:

# 2.1. Vocabulary size & richness by 3 markers: Type-Token Ratio, Brunet's Index, Honore's Statistic

# 2.2. Use count of the low-frequency words and the more specific words

# 2.3. Use count of noun-to-verb ratio (changes), low-image verbs, filler words, closed class words

# 2.4. POS tagging files and draw the marker-word-use count by year:

#    noun rate, pronoun rate, verb rate

# 2.5. Come up with a way for detecting human ability of verbal morphology

# 3. Syntactic Analysis:

# 3.1. # of clause per utterance.

# 3.2. Frequency of the use of passive voice

**Text Indexing techniques**

An inverted index is an index structure that maintains two hash indexed tables:-

1. Document\_table which is a set of document records containing two fields-
   * + 1. Doc\_id
       2. Posting\_list- list of terms in each document sorted according to some relevant measures.
2. Term\_table which is a list of terms containing two fields-
3. Term\_id
4. Posting \_list- list of document identifiers in which the term appears.

Such an index table will help us to easily answer our queries like “Find all documents associated with a particular term and so on.

**Text mining approaches-**

1. Keyword based analysis- collects sets of keywords or terms that occur frequently together and then finds the association or correlation relationship among them. Steps include- parsing, stemming, removing stop words, and so on and then evokes association mining algorithms.
2. Document classification analysis- Feature selection refers to removing those terms from the documents which are statistically not co-related. This will reduce the set of terms used for classification. After feature selection we can perform the classification on the ‘cleaned data’. The classification algorithms used-
3. Bayesian classification
4. Support vector machines.
5. Association based classification.

**Some other features**:-

(Empty speech in Alzheimer’s disease and fluent aphasia)

1. Empty phrases (contributing no content to the discourse)

2. Indefinite terms (thing, something, stuff)

3. Deictic terms (this ,that, here, there)

4. Pronouns without antecedent

5. Semantic par aphasia

6. Repeated words or phrases

7. Circumlocution (telling too many words which can otherwise be said in a much more shorter manner)

8. Neologisms- words with no apparent relation to a target eg: flakers for scissors)

9. Conjunctions- All conjunction excluding ‘and’ would render the discourse less coherent.

**Syntactic complexity**

Type 1: Length of production unit

1. Mean length of clause MLC # of words / # of clauses
2. Mean length of sentence MLS # of words / # of sentences
3. Mean length of T-unit MLT # of words / # of T-units

Type 2: Sentence complexity

1. Sentence complexity ratio C/S # of clauses / # of sentences

Type 3: Subordination

1. T-unit complexity ratio C/T # of clauses / # of T-units
2. Complex T-unit ratio CT/T # of complex T-units / # of T-units
3. Dependent clause ratio DC/C # of dependent clauses / # of clauses
4. Dependent clauses per T-unit DC/T # of dependent clauses / # of T-units

Type 4: Coordination

1. Coordinate phrases per clause CP/C # of coordinate phrases / # of clauses
2. Coordinate phrases per T-unit CP/T # of coordinate phrases / # of T-units
3. Sentence coordination ratio T/S # of T-units / # of sentences

Type 5: Particular structures

1. Complex nominals per clause CN/C # of complex nominals / # of clauses
2. Complex nominals per T-unit CN/T # of complex nominals / # of T-units Verb phrases per T-unit

Lexical

* + **Vocabulary size​**
  + Use of the **low-frequency words** and **more specific words**​.
  + **Noun rate**(reduce), both by type count and the word token count.
  + **Pronoun rate** (increase), both by type count and the word token count.
  + **Verb rate**, used to reflect the flow of the speech, both by type count and the word token count.
  + **Noun-to-Verb ratio** changes.
  + Use of **low-image verb**s.
  + **Lexical repetition**
  + **Filler words**, more use of filler words, more dysfluencies.
  + **Count of marker/labels (if any in transcripts)**: pause , hesitation, repeat, mumble, self-making/made-up words, substitution of generic terms for more specific ones.
  + **Adjective rate** and  **adverb rate** might be used but is not important or useful.
  + **Verbal morphology**. The deficits in this is regarded as the result of AD.
  + **Closed class words** use. These words contribute little to the content of texts but some believed that their use may be some function of a speaker's ability to form sentences.
  + **Type-Token Ratio** (**TTR**), is the ratio of total vocabulary V to overall text length N and is sensitive to the length of text collected.
  + ​**Brunet's Index W** is a length insensitive version of TTR: W = NV^(-0.165)
  + **Honore's Statistic R** is also insensitive to length: R = (100\*log N)/(1 - V1/V). V1 is the count of words only used once, the higher R value is, the  richer the vocabulary is.
* ​​​Syntactic  
  (This part is in fact still controversial. Quite a few papers and researchers show evidence that healthy aging contribute to the syntactic impairment, thus they don't think AD is the only reason even they believe AD is not the reason at all. Another issue is that the famous [1966 Sonwdon's Nun study](http://jamanetwork.com/journals/jama/article-abstract/396775) showed that the heterogenous grammar and syntactic ability will impact their performance in late life.   
    
  So far I have not seen any paper consider the heterogeneity in subjects' linguistic ability when performing the control experiment. The results comparison of the control experiments are not 'hypothesis-supported' said even by authors. I am working on to solve this issue now.)
  + The mean​ number of **clauses per utterance** (decrease).
  + The mean **length of sentences**.​
  + Decline in **grammatical complexity**.
  + Use of **passive voices** (eg. by phrases, be done phrases and get done phrases).
  + **Idea repetition** (by matching the cosine similarities of sentence vectors).
  + The mean **parse tree height/depth** of sentences (symmetric), (a measure of sentence embeddedness).
  + The mean, total, and maximum **Yngve depth** (an asymmetric measure that compensates for left-branching structures​), (a measure of sentence embeddedness).
  + **D-Level scale**, a psycholinguisticsbased ranking of sentence types into eight levels of increasing syntactic complexity​.

For the depression topic, I not long ago read a paper [Detecting late-life depression in Alzheimer's disease through analysis of speech and language-ACL 16’](http://www.aclweb.org/anthology/W/W16/W16-0301.pdf). This is the first time I know about it. This paper tries to use linguistic analysis and machine learning methods to distinguish AD and depression. But the achievement is not very good if you read it.

One high citation related paper -- [“Noncognitive” symptoms of early Alzheimer disease](https://www.researchgate.net/profile/Catherine_Roe/publication/270964544_Noncognitive_symptoms_of_early_Alzheimer_disease_A_longitudinal_analysis/links/5735dc4908ae298602e092de.pdf), this paper studies the depressive symptoms and other "non-cognitive" changes happen gradually before onset of cognitive impairment in AD dementia (which authors think being rarely studied). Some other research papers (e.g. [History of depression as a risk factor for Alzheimer's disease](http://pdfs.journals.lww.com/epidem/1995/07000/History_of_Depression_as_a_Risk_Factor_for.6.pdf)) show that the depression history of the patients, those who suffer from it from they were young occasionally or frequently, will have a double risk of ending up with AD.

Ftype token

According to [Alzheimer's association](https://www.alz.org/care/alzheimers-dementia-depression.asp), identifying depression in someone with Alzheimer's can be difficult, since dementia can cause some of the same symptoms. Examples of symptoms common to both depression and dementia include:

1. ​Apathy
2. ​Loss of interest in activities and hobbies
3. Social withdrawal
4. Isolation
5. Trouble concentrating
6. Impaired thinking

In addition, the cognitive impairment experienced by people with Alzheimer's often makes it difficult for them to articulate their sadness, hopelessness, guilt and other feelings associated with depression.

Depression in Alzheimer's doesn't always look like depression in people without Alzheimer's:

1. ​May be less severe
2. May not last as long and symptoms may come and go
3. The person with Alzheimer's may be less likely to talk about or attempt suicide.

There is no single test or questionnaire to detect depression. Diagnosis requires a thorough evaluation by a medical professional​ which is why the depression detection in AD patients is hard. I will pay attention to related research at the same time and see if  we can do something or define an interesting research question to solve.