DEMENTIA ANALYSIS THROUGH MACHINE LEARNING PART OF SPEECH TAGGERS

Abstract

The medical and technology fields have come together to perform astonishing feats in the past. Whether this be the development of the X-Ray, Covid-19 testing and tracking, or even remote surgeries, the technology and medical realms are intertwining. However, Machine Learning and the aspect of statistical modeling have just started to gain traction in medical practices. Only as recently as 2014 has machine learning passed its 1,000 publication mark in PubMed. As the art of Machine learning becomes more adept in the medicinal field, the task lends itself to early detection and treatment. It is, therefore, the aim of this study to detect and quantify language patterns in Dementia induces patients through the accessibility of Hidden Markov Models, Naive Bayes, etc.

HunPos is an open source implementation of the statistical part- of- speech tagger Trigrams'n Tags (TnT's) that allows the user to tune the tagger by using different feature settings. It is an early English adaptation used when branding part of speech taggers. In this paper, a reassessment of the HunPos tagger is done on WSJ articles; the pre-trained tool is then used to create a basis on Naive Bayes and other Part of Speech tagger operations trained on sample recordings from patients that suffer from Frontotemporal Lobe Degeneration (Mainly Dementia), and then an analysis is created on the specific features recorded w/ case studies.

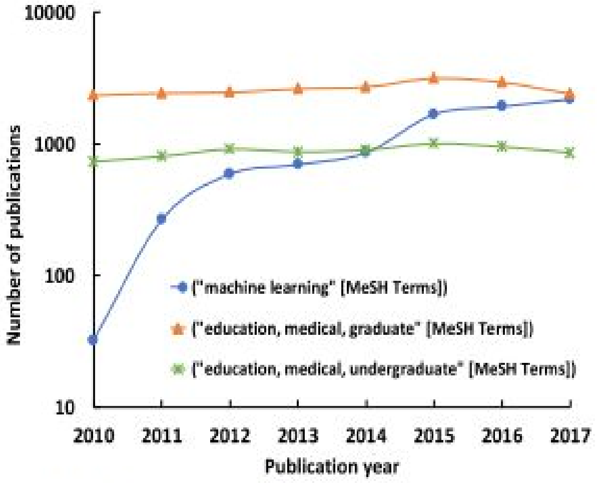


Fig. 1 Published papers within this decade as listed on US National Library of Medicine (PubMed) using "machine learning", "education, medical, graduate", and "education, medical, undergraduate" as MeSH terms, respectively. The actual user queries were: (i) "machine learning"[MeSH Terms] and ("2010/01/01"[PDAT]: "2017/12/31"[PDAT]), (ii) "education, medical, graduate"[MeSH Terms] and ("2010/01/01"[PDAT]: "2017/12/31"[PDAT]), and (iii) "education, medical, undergraduate"[MeSH Terms] and ("2010/01/01"[PDAT]: "2017/12/31"[PDAT])

KeyWords

Spontaneous speech, language, prosody, frontotemporal lobar degeneration, automated speech analysis, HunPOS, FLTD, POS tagging

1. Purpose

As a teenager who has gone through the struggles of having a speech impediment in the early stages of life, FLTD disorders hold a unique place in my field of interest. As speech impediments are more common within the younger population than the older population, the interest of this research came to incorporate the other side of the spectrum. Thus, the topic of dementia was picked.

Dementia is the general term that is associated with a loss of memory, language issues, problem solving abilities, and a range of other cognitive shortcomings. It falls underneath the FLTD category. FLTD is a heterogeneous disorder that comprises three different variants. These variants lead to the specific impact that dementia has on an individual. The more complex of a strain, the higher the discrepancy in the speech of the individual. Over half of the patients were diagnosed with FLTD exhibit language - related manifestations. The patients are then categorized as dysfluent, effortful, and agrammatical speech. However, all of these labels occur after a patient is already diagnosed with the disorder. The purpose of this research is to analyze different scenarios that can give hints on a person's FLTD status before the label is assigned.

2. Introduction

Dementia claims the persona of 10 million people every year around the world. It is the leading cause of death in England and Wales and is becoming a world leader in deaths overall. The only treatments are medications that are given at an age where the damage has already been done and where people's lives are at the brink. The alternative is a treatment that doctors can't recommend as the correct resources haven't emerged to their full potential yet. The treatment of early detection. To encourage such measures, a semi supervised machine learning algorithm model that introduces taggers to a subset of speech, known as part of speech tagging, is known to push the paradigms of medicinal practice.

3. Historical Background

Part of Speech tagging is defined as "the process of marking up a word in a text (also known as a corpus) as corresponding to a particular part of speech, based both on context and definition." The term corresponds with computational linguistics as it uses a set of algorithms to tag a subset of text. As will be mentioned, part of speech taggers fall underneath two categories; Rule based and shostatic. Rule based taggers include HunPos and E. Brill's taggers; both of which are used in this paper to analyze WSJ text. Both are plain English part of speech taggers but are overshadowed by the Universal Penn Treebank approach (Stanford NLP).

The Brill Tagger is an inductive method for POS tagging. It was created by Eric Brill in 1993 and is known to be an "error-driven transformation based tagger." This means that the tagger is assigned to a word based on the words frequency tag, but if it is unknown, it gets a "noun" tag. The tagger runs on two narratives:

- A form of supervised learning, which aims to minimize error
- A transformation-based process, in the sense that a tag is assigned to each word and changed using a set of predefined rules.

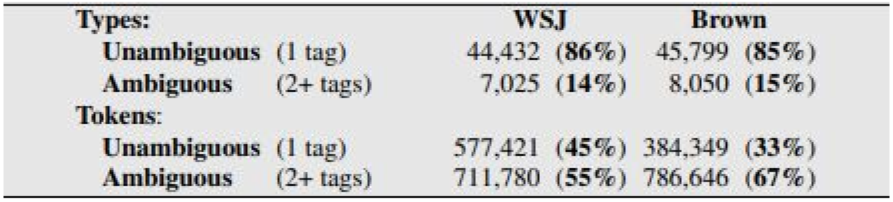
The text is first tokenized, so that each word can be analyzed individually, and then tagged based off of Lexical Rules. However, this model differs in one large way from the HMM models (that branch over the HunPos techniques). Here, rules are reapplied until a certain threshold is reached. Once this threshold is reached, no more rules can be put onto the constraint and the text annotations are halted.

tag1 → *tag2* IF *Condition*
The process for a Brill Tagger

\

On the other hand, we have the HMM taggers (Hidden Markov Models). These models are sequenced based. This means that the mapping of tags occurs in a sequence where the sequence is later captured in a sequence of labels. The HMM bases its reliance off of probability. After the model is given a sequence of units it calculates a probability distribution and picks the most suited one based on training data and test sets.

Like the HMM model, the Penn Treebank proves to produce very high accuracies. It has improved its accuracy from 95.3% from the 2010 decade to 97.3% token accuracy with a 56% sentence reading accuracy. However, as discussed by Manning (Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics?), the asymptote relies on specialized tokens and not whole sentence structure. As discussed, the only way to push the boundary is to take factors such as the gold standard, lexicon gaps, and unknown words into account. The gold standard, being the most inconsistent, lacks guidance. With more structured tokens, more ambiguous terms would be labeled correctly.

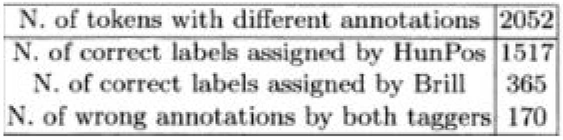


Types:		**WSJ**	**Brown**
Unambiguous	(1 tag)	44,432 (**86%**)	45,799 (**85%**)
Ambiguous	(2+ tags)	7,025 (**14%**)	8,050 (**15%**)
Tokens:			
Unambiguous	(1 tag)	577,421 (**45%**)	384,349 (**33%**)
Ambiguous	(2+ tags)	711,780 (**55%**)	786,646 (**67%**)

Tag Ambiguity for word types in Brown and WSJ using the Penn Treebank (45 tags) (Stanford POS Tagging).

3.1 Reassessment

For this paper, a trusted tagger must be used. One that can most accurately convey the English lexicon (after running through the WSJ corpora) and one that can be compared for cost analysis on different cases. Thus, an HMM tagger (HunPos) and the Brill Tagger were put on test for tokenized accuracy. The Data used came from NLTK Toolkit. Both taggers were already semi trained on smaller chunks of data, so the taggers had to be tested on the WSJ corpus.



N. of tokens with different annotations	2052
N. of correct labels assigned by HunPos	1517
N. of correct labels assigned by Brill	365
N. of wrong annotations by both taggers	170

Tagger Evaluation on WSJ

As shown, in a set of 10,493 tokens taken from the NLTK database (biology based texts) the HunPos tagger did better. As it may have done better on larger texts, it still performs around the same accuracy (74%) as the Brill tagger on smaller texts. Here, on larger scales, the HunPos tagger tags around 97%; around the same accuracy that the NLTK suite holds for the Brown corpus.

3.2 Related Tests

As shown by the taggers accuracy, the task at hand comes to tag dementia related scripts in a certain way to fixate a pattern. In March of 2020, researchers at Hindawi tested on 5,272 patients. They gave the patients a 37item questionnaire in which “Information Gain” came out as the most effective of the three feature spaces.

The task of the research was to categorize patients with NC, MCI, VMD, or dementia. NC stood for normal cognition (didn’t meet the NIA-AA criteria for Alzeimers), MCI for a cognitive disorder in the realms of orientation/judgement (with a CDR less than 0.5), and VMD for a CDR score above 0.5, mental disabilities in 2 or more domains and a mild decline in social daily activities. The researchers divided the 5,272 patients into two categories; 4,745 patients as a training group and 527 patients as the test group. In the training set, there were 328 for normal, 1,234 for MCI, 718 for VMD, and 2,465 for dementia. In the test set, there were 51 for normal, 113 for MCI, 98 for VMD, and 265 for dementia.

As mentioned before, the research held high standards for three feature selection methods that improved the generalization for the models.

- The Random Forest Algorithm for Feature Selection.
  - This algorithm randomly assigns weights to paths in a top - down method. After continuous testing an out-of-bag error rate is retained and the method with the lowest error rate is used at the end of reiteration.
- The Information Gain for Feature Selection
  - The model relies on how much information the feature can bring to the model as a whole. The idea is based on entropy; used to evaluate importance of a certain aspect in a feature.

Shannon's entropy equation:

$$H(X) = -\sum_{i=0}^{N-1} p_i \log_2 p_i$$

- The Relief Algorithm for Feature Selection
  - The algorithm relies on neighbor branches having similar weight values while different classes of nearest neighbors differ greatly. This is a very efficient method but does not account for redundancy.

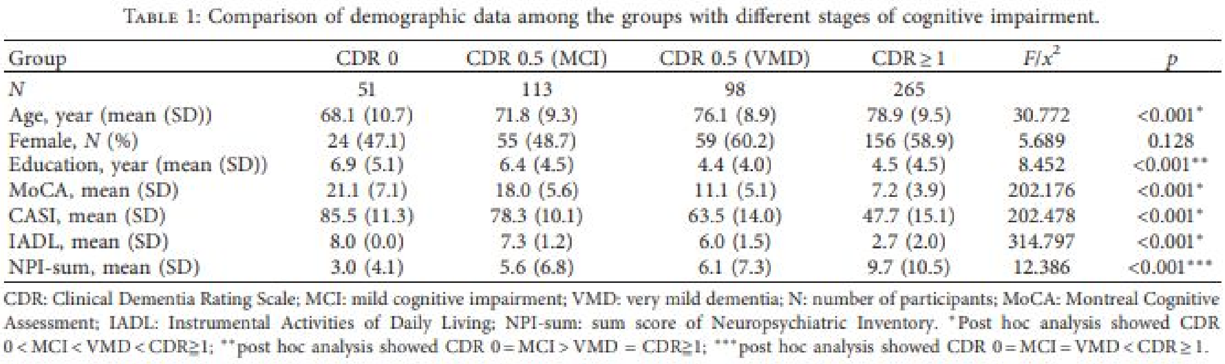
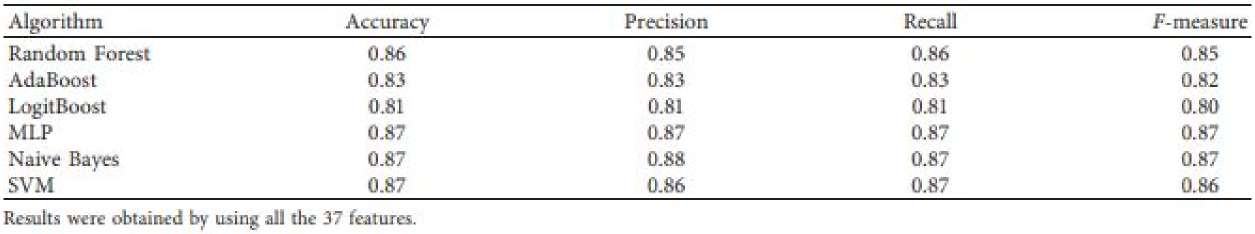


TABLE 1: Comparison of demographic data among the groups with different stages of cognitive impairment.

Group	CDR 0	CDR 0.5 (MCI)	CDR 0.5 (VMD)	CDR ≥ 1	F/x^2	p
N	51	113	98	265		
Age, year (mean (SD))	68.1 (10.7)	71.8 (9.3)	76.1 (8.9)	78.9 (9.5)	30.772	<0.001*
Female, N (%)	24 (47.1)	55 (48.7)	59 (60.2)	156 (58.9)	5.689	0.128
Education, year (mean (SD))	6.9 (5.1)	6.4 (4.5)	4.4 (4.0)	4.5 (4.5)	8.452	<0.001**
MoCA, mean (SD)	21.1 (7.1)	18.0 (5.6)	11.1 (5.1)	7.2 (3.9)	202.176	<0.001*
CASI, mean (SD)	85.5 (11.3)	78.3 (10.1)	63.5 (14.0)	47.7 (15.1)	202.478	<0.001*
IADL, mean (SD)	8.0 (0.0)	7.3 (1.2)	6.0 (1.5)	2.7 (2.0)	314.797	<0.001*
NPI-sum, mean (SD)	3.0 (4.1)	5.6 (6.8)	6.1 (7.3)	9.7 (10.5)	12.386	<0.001***

CDR: Clinical Dementia Rating Scale; MCI: mild cognitive impairment; VMD: very mild dementia; N: number of participants; MoCA: Montreal Cognitive Assessment; IADL: Instrumental Activities of Daily Living; NPI-sum: sum score of Neuropsychiatric Inventory. *Post hoc analysis showed CDR 0 < MCI < VMD < CDR≥1; **post hoc analysis showed CDR 0 = MCI > VMD = CDR≥1; ***post hoc analysis showed CDR 0 = MCI = VMD < CDR ≥ 1.

The table above represents the remaining 527 patients and their dementia classification.



Algorithm	Accuracy	Precision	Recall	*F*-measure
Random Forest	0.86	0.85	0.86	0.85
AdaBoost	0.83	0.83	0.83	0.82
LogitBoost	0.81	0.81	0.81	0.80
MLP	0.87	0.87	0.87	0.87
Naive Bayes	0.87	0.88	0.87	0.87
SVM	0.87	0.86	0.87	0.86

Results were obtained by using all the 37 features.

The Naive Bayes model worked the most efficiently. The algorithm improved normal sensitivity by .1, MCI by .31, VMD by .21, and dementia by .03.

The conclusions drawn from this are that Information gain worked best with the Naive Bayes model to improve the accuracy of detecting early dementia. Unlike HunPos, a Trigram tagger set that is modeled on a set of text, Naive Bayes models are probability based models that push the accuracy of a feature set.

The accuracy labeled on HunPos after being run through a WSJ is primarily used for text based operations while the Naive Bayes method draws distinct correlations in a large data set and has room for improvement.

Continued...

More like HunPos taggers comes a POS tagger experiment that was conducted to analyze the speech portion of dementia stricken patients (Computerized Analysis of Speech and Language to Identify Psycholinguistic Correlates of Frontotemporal Lobar Degeneration).

Language related manifestations are diverse based on the individual and are currently classified through 2 different approaches.

- The first involves subjective assessment that are dictated by trained neuropsychologists
  - The dimensions usually tested here are speaking rate, distortion of sound, and grammar mistakes)
- The second approach shifts its focus to machine learning algorithms. It includes phonologic, syntactic, semantic, and pragmatic features that are based off of text. This gives a more objective sense to the situation.

The test was conducted through a group of "Thirty-eight patients diagnosed with 1 of the 3 FTLD syndromes (bvFTD, PNFA, SD) and PLA were recruited from 3 academic medical centers. All aspects of this study were approved by the IRBs at each of the medical centers and the University of Minnesota. All 38 participants underwent a neuropsychologic test battery that included the Boston Diagnostic Aphasia Examination Cookie-Theft Picture Description Task." The task defined peoples recordings/text into 4 categories; bvFTD, PNFA, PLA, and SD.

As stated by the article:

"PNFA was diagnosed with expressive speech characterized by at least 3 of these: nonfluency (reduced numbers of words per utterance), speech hesitancy or labored speech, word finding difficulty, or agrammatism, in which these symptoms constitute the principal deficits and the initial presentation.

PLA was diagnosed with fluent aphasia with anomia but intact word meaning and object recognition in which these symptoms constitute the principal deficits and the initial presentation.

SD was diagnosed with loss of comprehension of word meaning, object identity, or face identity in which these symptoms constitute the principal deficits and the initial presentation.

Behavioral variant FTD was diagnosed with a change in personality and behavior sufficient to interfere with work or interpersonal relationships; these symptoms constituted the principal deficits and the initial presentation, and had at least 5 core symptoms in the domains of aberrant personal conduct and impaired interpersonal relationships."

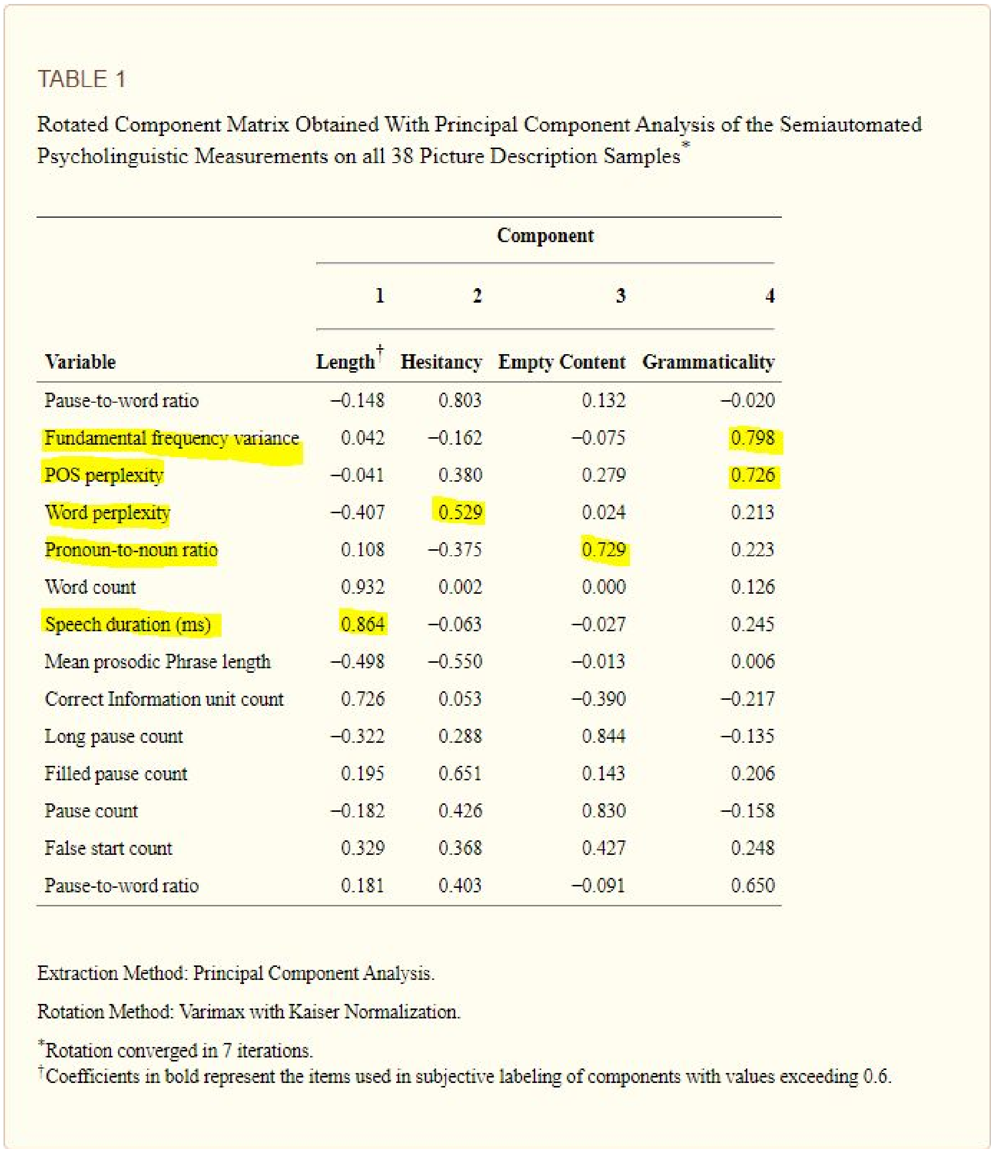


TABLE 1

Rotated Component Matrix Obtained With Principal Component Analysis of the Semiautomated Psycholinguistic Measurements on all 38 Picture Description Samples*

	Component			
	1	2	3	4
Variable	Length†	Hesitancy	Empty Content	Grammaticality
Pause-to-word ratio	−0.148	0.803	0.132	−0.020
Fundamental frequency variance	0.042	−0.162	−0.075	0.798
POS perplexity	−0.041	0.380	0.279	0.726
Word perplexity	−0.407	0.529	0.024	0.213
Pronoun-to-noun ratio	0.108	−0.375	0.729	0.223
Word count	0.932	0.002	0.000	0.126
Speech duration (ms)	0.864	−0.063	−0.027	0.245
Mean prosodic Phrase length	−0.498	−0.550	−0.013	0.006
Correct Information unit count	0.726	0.053	−0.390	−0.217
Long pause count	−0.322	0.288	0.844	−0.135
Filled pause count	0.195	0.651	0.143	0.206
Pause count	−0.182	0.426	0.830	−0.158
False start count	0.329	0.368	0.427	0.248
Pause-to-word ratio	0.181	0.403	−0.091	0.650

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

*Rotation converged in 7 iterations.
†Coefficients in bold represent the items used in subjective labeling of components with values exceeding 0.6.

Through the means of ASR (the method to turn audio into text) and NLP-POS methods, the following results were taken.

Ratios of around .6 on either side of 0 shows intentions of severity and can be used as a baseline for early development. As highlighted, POS taggers initialized the means of audio and text that are the fundamental differences patients with FLTD undergo.

The specific corpus used in the study was one used from spontaneous telephone calls (SWITCHBOARD). SWITCHBOARD contained around 14,580 minutes of calls that were transposed by a TnT part of speech tagger (HunPos). The tagger calculated the part of speech basis off of the words around it (Brant, 2001).

4. Case Theory

The following are two different case reports cited from the BMJ Journal.

Patient 1: Dementia

87 Year old Man with dementia and Lewy bodies. Rapid functional decline and weight loss have been associated with injuries over the last 9 months. The man "satisfied both the 2005 and 2017 criteria for the diagnosis of probable DLB including the onset of dementia prior to motor symptoms and the presence of two or more core features: fluctuating cognition and alertness, well-formed visual hallucinations and spontaneous parkinsonism features.

The medical journal concludes:

"A texture-modified diet had been implemented due to concerns surrounding dysphagia and potential aspiration. The patient was observed to rapidly deteriorate in health status, precipitated by two injurious falls over 9 months while enrolled in a wait-list control period for the exercise trial. Evaluations were conducted to identify potential aetiologic factors in his rapid functional decline" [Can condense with my own words a bit more].

Patient 2: Early Dementia symptoms

56 year old man who has been losing parts of his memory over the past ten years. He had engaged in mood swings, social isolation, and slurring of his words. The patient was reported to forget his own family members names many of the times, forget where he was at times, and used to get less than 3 hours of sleep per night.An increased risk of anxiety and agitation was present at times. The patient had an MRI taken and an EEG performed. Both came back normal. Carbamazepine and quetiapine were added to the patients diet but didn't seem to haven a substantial effect.

The medical journal concludes:

"A mental status examination revealed average grooming and hygiene, poor eye-to-eye contact, psychomotor retardation, a calm and quiet demeanour, irrelevant yet coherent responses to questions and blunted affect. The patient denied experiencing any hallucinations or other perceptual abnormalities and no overt delusions or obsessions were noted during his examination.

A cognitive assessment with the mini-mental state examination revealed a score of zero, as the patient was unable to complete any of the required tasks. While the patient registered three items correctly, he recalled none. He was unable to do the serial sevens or three measures of attention and was unable to read or write."

The patient had not completed school past secondary school and worked as an electrician for 20 years due to personal issues. No other abnormalities were taken from the test.

5. Analysis
[To be written on 6/29/20]

6. Sources

1. Elhusein, B., Mahgoub, O., & Khairi, A. (2020, March 01). Early-onset dementia: Diagnostic challenges. Retrieved June 28, 2020, from https://casereports.bmj.com/content/13/3/e233460

2. Inskip, M., Mavros, Y., Sachdev, P., & Singh, M. (2020, April 01). Interrupting the trajectory of frailty in dementia with Lewy bodies with anabolic exercise, dietary intervention and deprescribing of hazardous medications. Retrieved June 28, 2020, from https://casereports.bmj.com/content/13/4/e231336

3. Klabunde, R. (2002). Daniel Jurafsky/James H. Martin, Speech and Language Processing. *Zeitschrift Für Sprachwissenschaft, 21*(1). doi:10.1515/zfsw.2002.21.1.134

4. Kolachalama, V. B., & Garg, P. S. (2018). Machine learning and medical education. *Npj Digital Medicine, 1*(1). doi:10.1038/s41746-018-0061-1

5. Manning, C. D. (2011). Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics? *Computational Linguistics and Intelligent Text Processing Lecture Notes in Computer Science,* 171-189. doi:10.1007/978-3-642-19400-9_14

6. Martinez, A. R. (2011). Part-of-speech tagging. *Wiley Interdisciplinary Reviews: Computational Statistics, 4*(1), 107-113. doi:10.1002/wics.195

7. Pakhomov, S., Smith, G., Chacon, D., Feliciano, Y., Graff-Radford, N., Caselli, R., & Knopman, D. (2010, September). Computerized analysis of speech and language to identify psycholinguistic correlates of frontotemporal lobar degeneration. Retrieved June 28, 2020, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3365864/

8. Proceedings IWPC 2000. 8th International Workshop on Program Comprehension. (2000). *Proceedings IWPC 2000. 8th International Workshop on Program Comprehension*. doi:10.1109/wpc.2000.852473

9. Sharp, B. (n.d.). Human-Machine Interaction in Translation: Proceedings of the 8th International NLPCS Workshop. Retrieved June 28, 2020, from https://books.google.com/books?id=jDpS6D60o8AC

10. Zhu, F., Li, X., Tang, H., He, Z., Zhang, C., Hung, G., . . . Zhou, W. (2020). Machine Learning for the Preliminary Diagnosis of Dementia. *Scientific Programming, 2020*, 1-10. doi:10.1155/2020/5629090