2019 16th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE) Mexico City, Mexico. September 11-13, 2019

Search for Dementia Patterns in Transcribed Conversations using Natural Language Processing

Damián Solís Rosas
Facultad de Ingeniería
Universidad Autónoma de Querétaro
Querétaro, México
damian.solis@hotmail.com

Saúl Tovar Arriaga
Facultad de Ingeniería
Universidad Autónoma de Querétaro
Querétaro, México
saul.tovar@uaq.mx

Marco Antonio Aceves Fernández
Facultad de Ingeniería
Universidad Autónoma de Querétaro
Querétaro, México
marco.aceves@gmail.com

***Abstract*—The effects on the linguistic capacity of the people with some type of dementia are reflected in the lexicon (his mental dictionaries and his ability to understand complex words) rather than his ability to formulate complete and fluent enunciations. Analysis indicate that the richness of the lexicon and the fluency to speak are not good qualities in people who suffer Alzheimer. There are previous studies of the pathologies of speech, which include the use of pauses, words of filling, words formulated, restarts, repetitions, incomplete statements and diffluent speech. All previous factors may occur in individuals with some type of dementia. Through discriminatory analysis of conversation and metrics analysis, we found slight statistical differences between people with and without dementia. Additionally, we use two machine learning algorithms to automatically classify presence/absence of dementia. The first one, a 3-layer neural network reaching a binary classification accuracy of 78.3%, and the second a support vector machine reaching a binary classification accuracy of 86.42%.**

***Index Terms*—Natural Language Processing, Dementia Patterns, Text Classification, Support Vector Machine, Neural Networks**

I. Introduction

Dementia is a general term for a decline in mental ability, describes a set of symptoms that may include memory loss and difficulties with thinking, problem-solving or language. A person with dementia may also experience changes in their mood or behaviour to such an extent that it interferes with a person's daily life and activities.

Most patients with some type of dementia are characterized by the degradation of their language and cognitive functions resulting in significant communication difficulties.

The detection of dementia can be an expensive and exhausting process for the person to be diagnosed [1].

Based in how far a person's dementia has progressed, dementia can be divided in stages. Defining a person's disease stage helps physicians determine the best treatment approach and aids communication between health providers and caregivers. The stages of Dementia can be grouped in Early-Stage (Moderate Cognitive Decline), Mid-Stage (Severe Cognitive Decline) and Late-Stage (Very Severe Cognitive Decline).

Each Stage has signs and symptoms. In the Early-Stage people can quickly forget what they have heard, seen or thought, it is common for them to lose the topic and repeat an idea many times, consequently, it is difficult to follow a conversation.

In the Mid-Stage people are often disoriented in time and space, usually have large memory deficits, so they can not remember recent events, these characteristics generate a poor verbal production, meaningless and simple.

In the Late-Stage people have serious problems to pay attention, to codify, retrieve information, have perception problems and their executive functions are limited or lack them. Semantic memory and the ability to remember concepts are very deteriorated or seem absent in the person.

In all the stages of dementia, the characteristics and linguistic capacities are deteriorated, these characteristics can serve as object of study and analysis to get an early diagnosis.

It is possible to find markers that give signs of early dementia by analyzing and processing the natural language obtained from patients through medical tests, questionnaires or simple conversations [2].For the analysis and search of markers in conversations, it is necessary to know the characteristics of communication in people who suffer dementia. We can found these characteristics from the Early-Stage.

The communication characteristics of people with dementia vary in each stage. As dementia progresses, communication problems increase, causing a deterioration in people's communication skills. The deterioration in communication impacts in daily life of patients and deteriorates one of the main tools and qualities of the human being: communication. Losing the ability to communicate can be one of the most frustrating and difficult problems for people with dementia, their families and carers. They find it more and more difficult to express themselves clearly and to understand what others say.

The main characteristics of the communication of people according to the stage of dementia are [3]:

- Early-Stage: Difficulty to understand sentences with complex content. Conserved syntactic structure and problems

978-1-7281-4840-3/19/$31.00 © 2019 IEEE

to repeat long sentences. Phonological system conserved. Incomplete Ideas.
- Mid-Stage: Difficulty to nominate and categorize. Reduced expressive vocabulary. Difficulty to repeat simple sentences. Omission of connectors and functional words in the sentences. Incomplete Phrases.
- Late-Stage: Reduction in Vocabulary. Frequent omission of functional words. Uncontrolled repetition of phrases said by the speaker. Uncontrolled repetition of the same word.

The conversations can be classified applying an statistical analysis and computational techniques such as support vector machine, decision tree and Bayes classifier[1][2][3][4].

In the next table are shown the previous studies with their authors, objective and results.

TABLE I
STATE OF THE ART

Author	Data Format	Technique	Accuracy
C. Guinn Ben Singer A. Habash [1]	Transcribed Conversations	Decision Tree Bayesian	67% 80
A. Khodabakhsh S. Kuscuoglu C. Demiroglu [2]	Transcribed Conversations	Decision Tree SVM	90% 80%
B. Roark M. Mitchell J. Hosom K. Hollingshead [3]	Audio	SVM	86%
C. Thomas M. Mitchell Vlado Keselj Kenneth Rockwood [4]	Transcribed Conversations	Bayesian	90%

A neural network can be used to classify text, the accuracy obtained by the neural network can be compared with other classification methods to evaluate its performance.

II. Materials and Methods

A. Statistical Analysis

To process and analyze the text, the NLP (Natural Language Processing) use techniques as Lemmatization, Morphological segmentation, Word segmentation, Stemming, etc.

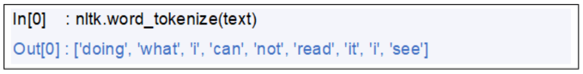
Based on the lexical and grammatical characteristics of people who have some type of dementia and those who do not have it and with the help of computational techniques, a computer system can be developed to find markers in the conversations of these people. For the analysis of the conversations and the application of techniques, algorithms were implemented with the help of the NLTK (Natural Language Toolkit) [4] library.

In this study, to get markers in the conversations, conversations of people who have some type of dementia and people who do not have it were used. These conversations are in the English language and are part of the Carolina Corpus Conversation database[5].

The Carolina Corpus Conversation database has hundreds of conversations of people with some pathological condition, the conversations are transcribed and include features such as: Comments from the recorder; Sounds recorded during conversations; Feelings reflected by the people involved in the conversations. Pauses, reflected feelings and signs of bewilderment or confusion are represented in conversations by different signs or chains of signs.

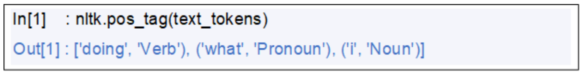
The techniques that were used to search characteristics in the conversations were:

- Text tokenization.



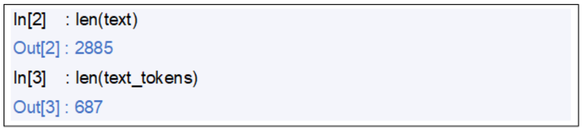
```
In[0]   : nltk.word_tokenize(text)
Out[0] : ['doing', 'what', 'i', 'can', 'not', 'read', 'it', 'i', 'see']
```

- Part of Speech



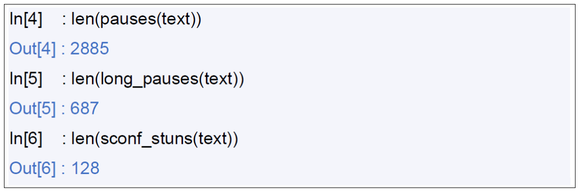
```
In[1]   : nltk.pos_tag(text_tokens)
Out[1] : ['doing', 'Verb'), ('what', 'Pronoun'), ('i', 'Noun')]
```

- Word Count.



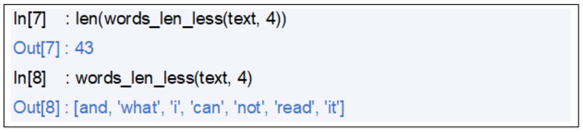
```
In[2]   : len(text)
Out[2] : 2885
In[3]   : len(text_tokens)
Out[3] : 687
```

- Count pauses, long pauses and stuns or confusions recorded in the conversation.



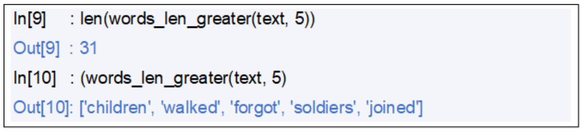
```
In[4]   : len(pauses(text))
Out[4] : 2885
In[5]   : len(long_pauses(text))
Out[5] : 687
In[6]   : len(sconf_stuns(text))
Out[6] : 128
```

- Word count of length less than or equal to 4



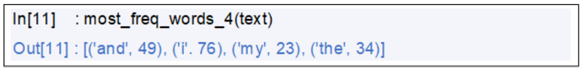
```
In[7]   : len(words_len_less(text, 4))
Out[7] : 43
In[8]   : words_len_less(text, 4)
Out[8] : [and, 'what', 'i', 'can', 'not', 'read', 'it']
```

- Word count of length greater than or equal to 5



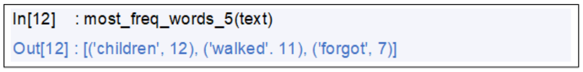
```
In[9]    : len(words_len_greater(text, 5))
Out[9]  : 31
In[10]  : (words_len_greater(text, 5)
Out[10]: ['children', 'walked', 'forgot', 'soldiers', 'joined']
```

- Extraction of the most used words of length less than or equal to 4 and frequency



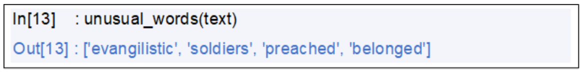
```
In[11]   : most_freq_words_4(text)
Out[11] : [('and', 49), ('i'. 76), ('my', 23), ('the', 34)]
```

- Extraction of the most used words of length greater than or equal to 5 and frequency



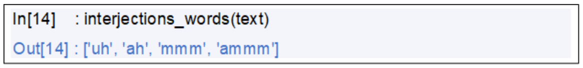
```
In[12]   : most_freq_words_5(text)
Out[12] : [('children', 12), ('walked'. 11), ('forgot', 7)]
```

- Counting and extraction of unusual words



```
In[13]   : unusual_words(text)
Out[13] : ['evangilistic', 'soldiers', 'preached', 'belonged']
```

- Counting and extraction of interjections



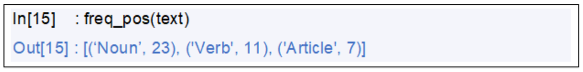
```
In[14]   : interjections_words(text)
Out[14] : ['uh', 'ah', 'mmm', 'ammm']
```

TABLE II

CHARACTERISTICS OF COMMUNICATION ACCORDING TO THE STAGE OF DEMENTIA

	Semantic	**Sintactic**	**Phonological**	**Pragmatic**	**Literacy**
Early-Stage	Difficulty to understand sentences with complex content **N.A.**	Conserved syntactic structure and problems to repeat long sentences **Possible**.	Phonological system conserved **N.A.**	Ramble **Possible (A.I.)** Incomplete Ideas **Possible (A.I.)**	Disortography **N.A.**
Mid-Stage	Difficulty to nominate and categorize **Possible**. Reduced expressive vocabulary **Possible**.	Difficulty to repeat simple sentences **Possible**. Omission of connectors and functional words in the sentences **Possible**.	Occasional confusion of pronunciation patterns **N.A.**	Incomplete Phrases **Possible (A.I.)** Repetition of ideas in the conversation **Possible (A.I.)** Loss of topic and leave the conversation **Possible (A.I.)**	Writing of words and short phrases **N.A.**
Late-Stage	Reduction in Vocabulary **Possible**. Use only significant elements **Possible (A.I.)** Semantic paraphasias **Possible (A.I.)**	Limited automatic language **Possible (A.I.)** Almost zero repetition, even for monosyllabic words **Possible**. Frequent omission of functional words **Possible**.	Uncontrolled repetition of phrases said by the speaker **Possible**. Uncontrolled repetition of the same word **Possible**. Uncontrolled repetition of syllables **Possible**.	Conversation almost absent, limited or inaccurate **Possible**. Impossibility to maintain the topic **Possible (A.I.)**	Almost total impairment of writing **N.A.**

- Counting, extraction and analysis of Part of Speech



```
In[15]   : freq_pos(text)
Out[15] : [('Noun', 23), ('Verb', 11), ('Article', 7)]
```

To use the techniques mentioned previously, the conversations were preprocessed; punctuation marks were removed; the pauses and recorded stuns were coded; all text was converted to lowercase; The names of the people involved were eliminated. All the above was done for the counting and extraction of characteristic data that were later used to obtain statistical results and use artificial intelligence techniques.

Table II shows the communication characteristics in each stage of dementia. After an analysis of the database, taking into consideration the nature of the questions and answers, we labeled each of the characteristics found in each dementia stages. The first category is the N.A.(Not Available), which means that taking into consideration the available information in the database, to our knowledge, it is not possible to get the pattern. E.g., we can not find if a person have difficulties to understand sentences. Possible means that these characteristics can be analyzed using statistical techniques, and Possible A.I. means that this characteristics may be analyzed using Artificial Intelligence techniques.

The lexical diversity (LD) tells us how much vocabulary was deployed. This measure is important because the vocabulary of people with dementia in Mid-Stage and Late-Stage is very poor (Table II). The greater value of lexical diversity means that more different words were used in the conversation. The lexical diversity is obtained by dividing the number of words in the text Text Length (TL) by the number of words of the vocabulary used Vocabulary Length (VL). The formula is:

$$LD = TL/VL \tag{1}$$

The size of the words is a characteristic with which it can be inferred that a word has greater importance in the text, since the words of small length are usually articles, pronouns, conjunctions, etc. Persons with dementia tend to decrease using long words and replace them by simple ones easier to remember. Analyzing the frequency of long length words, we can infer what is the main topic or the topics that were discussed in the conversation.

Unusual words are those that are not present in informal language and even some of them are not in dictionaries. Some of them are scientific jargon and others are even invented by people. Persons with dementia tend to use these 'invented' words without noticing that they do not belong to the language. The percentage of unusual words in the vocabulary were counted and calculated. To obtain the percentage of unusual words (PPL), the number of unusual words (NPL) is divided by the number of words of the vocabulary used Vocabulary Length (VL) and multiplied by 100 %. The formula is:

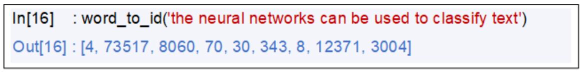
$$PPL = (NPL/VL) * 100\% \tag{2}$$

The number of interjections, pauses and stun or confusion were counted. A high percentage of those words are a sign of Pragmatic issues (Table II).

An analysis of the words used was made based in the Part of Speech. The amount of articles, pronouns, conjunctions and adverbs used with respect to nouns and verbs give us a statistic of the complexity of the sentences formed in conversations [1].

B. Automatic Classifiers

An additional way to do an automatic classification of persons with and without dementia is using a neural network (NN). The application of the neural network in this case is to make a binary classification. To train the neural network, the conversations were encoded because we can not use raw text in the networks. Each conversation was divided in blocks of 256 words, a number was assigned to each word according to a dictionary. The data used in the neural networks are arrays of numbers with length of 256.
For example, the sentence *'The neural networks can be used to classify text'* is encoded:



```
In[16]   : word_to_id('the neural networks can be used to classify text')
Out[16] : [4, 73517, 8060, 70, 30, 343, 8, 12371, 3004]
```

The NN is a sequential model. The Sequential model is a linear stack of layers. This NN has 3 layers, the first and the second have 16 nodes and every node implements a rectified linear activation unit, the third has 1 node and this one implements a Sigmoid function.

The NN use the Adam optimizer, this is an optimization algorithm that can used instead of the classical stochastic gradient descent procedure to update network weights iterative based in training data. The loss function is measured with the Binary Cross-Entropy, wich is a loss function used on problems involving binary decisions.

The software used to make the automatic classifier is TensorFlow[6] combined with Keras[7].

To train the neural network the data was obtained from 62 conversations from people who have dementia and 160 from people who do not have dementia. Each conversation was tagged with 1 if the person have dementia and 0 if the person do not have dementia.

Another way to do an automatic classification of persons with and without dementia is using a support vector machine (support vector machine). Like the neuronal network the SVM is applied to make a binary classification. To apply the SVM the stop words were removed; the stem of the words were subtracted; the words obtained after the first steps were divided in four groups: nouns, adjectives, verbs and adverbs.

After processing the words, the conversations were transformed in vectors then the SVM were applied.

The software used to apply the SVM is Scikit-learn[8].

III. Results

The results showed below are from conversations of 20 persons, 10 of them have dementia and 10 do not have dementia. The lexical diversity, the percentage of words with length less than or equal to 4 and the percentage of words with a length greater than or equal to 5 are shown in the Table III. The persons who have dementia are represented with the tag ”CD” and who do not have dementia with the tag ”SD”.

TABLE III
Lexical Diversity, Used Words and Interjections in the conversations

Person	Lexical Diversity	Words l <=4 (%)	Words l >=5 (%)	Interjections (%)
CD	2.03	64.14	31.81	4.05
CD	2.08	62.99	34.24	2.77
CD	1.79	68.95	28.15	2.9
CD	1.99	67.54	29.97	2.52
CD	2.47	74.38	22.92	2.7
CD	2.31	69.11	28.77	2.12
CD	2.04	71.32	26.67	2.01
CD	2.21	65.87	31.10	3.03
CD	2.48	62.43	34.83	2.74
CD	1.85	59.42	35.61	4.97
Average	2.12	66.61	33.39	2.98
SD	2.51	66.54	32.29	1.17
SD	1.87	70.27	27.04	2.69
SD	2.37	67.31	30.29	2.4
SD	2.28	69.81	27.48	2.71
SD	2.53	68.42	29.31	2.27
SD	1.98	65.41	32.91	1.68
SD	2.51	72.32	26.64	1.04
SD	2.34	71.29	25.62	3.09
SD	2.41	69.31	29.81	0.88
SD	2.07	70.51	28.86	0.63
Average	2.28	69.11	30.89	1.85

The following table shows the percentage of daze or confusions, short pauses, long pauses and unusual words recorded in conversations.

TABLE IV
Daze or confusion, short pauses, long pauses and unusual words recorded in the conversations

Person	Daze/ Confusion	Short Pauses (%)	Long Pauses (%)	Unusual Words (%)
CD	2.78	3.18	1.11	6.01
CD	1.42	1.17	0.76	8.14
CD	1.2	2.87	1.05	8.05
CD	0.98	2.34	0.97	7.3
CD	1.13	1.83	0.81	9.76
CD	0.87	2.94	1.21	8.47
CD	2.3	4.12	2.03	10.51
CD	1.92	3.35	1.52	10.26
CD	0.76	2.1	0.84	9.11
CD	0.81	1.87	0.73	7.63
Average	1.41	2.57	1.1	8.52
SD	0	0.56	0	1.97
SD	0	0.75	0	6.19
SD	0	1.9	0	5.77
SD	0.87	0.46	0.3	4.21
SD	0.76	1.37	0.41	7.88
SD	0	1.13	0.51	2.04
SD	0	0.87	0.21	3.1
SD	0.3	1.26	0	2.91
SD	0	1.1	0	4.51
SD	0	1.35	0	3.77
Average	0.19	1.07	0.14	4.23

The percentages indicate how much the words were used throughout the conversation. The row of average shows the averages of each metric of the words according to the classification of the conversations.

An analysis of the Part of Speech was made using techniques of the Natural Language Toolkit and statistical techniques. In Figure 1 are shown the main types of word which are conjunctions, adjectives, determiners, and some others in a Box Chart. The blue chart represent the Part of Speech of people who have dementia and the gray one represent the Part of Speech of people who do not have dementia.

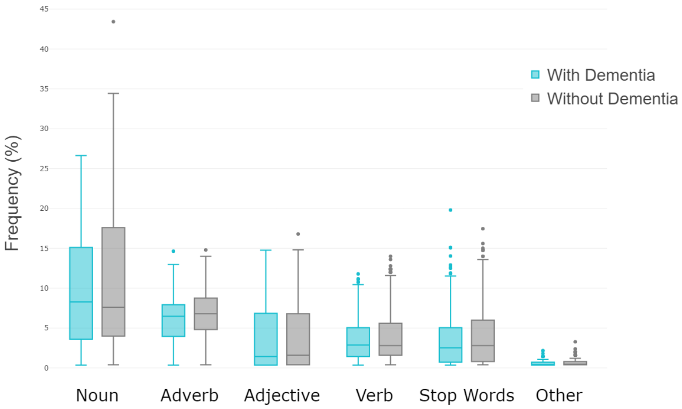


Fig. 1 - Types of words used in the conversations

There are a visual difference in the groups of noun, adverb, determiner, conjunction and preposition in the people who have dementia with the people who do not have it.

A 3-layer neural network was used, with 25 epochs, after the twenty-five epoch the neural network suffer an overfitting. To train the neural network 70% of the data was used and the remaining 30% was used to validate. The training and validation accuracy of the neural network are shown in the Figure 2.

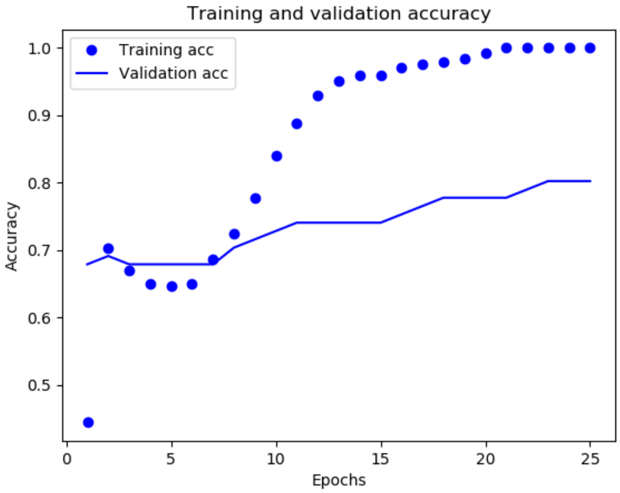


Fig. 2 - Training and validation accuracy of the NN

As we can see in Figure 2 the accuracy is improved in each epoch.

To apply the support vector machine classifier 70% of the data was used to training and the remaining 30% was used to validate.The accuracy of the Support Vector Machine to classify the conversations was 86.42%.

IV. Conclusion

According to our measures of lexical diversity, used words and interjections we can not find a significant or conclusive pattern to determine is a person has dementia. With statistical analysis we can not found a conclusive pattern in the conversations. There are a lot of factors that are not represented in the data such as: the stage of dementia and the education level [9]. According to previous studies the stage of dementia and the education level are directly related with the communication.

In contrast, our results in percentage of daze, pauses and unusual words show that it is useful to take this metric into consideration. The results indicate that these metrics are greater in the people who have dementia, because the capacities to understand, reason, learn and enunciate words are deteriorated in people with dementia according to the stage in which they are.

As we can see in the figure 1, the groups of noun, adverb, determiner, conjunction and preposition have a visual difference, in people who have dementia are fewer used, as we said previously. The language in people who have dementia is simple, the sentences that they use have a lack of connectors, descriptors and nouns they want to say something in the easiest way.

The neural network used obtained a score of 78.3% in the classification of conversations of people with dementia and without dementia. Different networks were trained and the best score was obtained with a 3-layer neural network in 25 epochs. This accuracy may be improved using data of more subjects or adding data such as stage of dementia and the education level.

The support vector machined used obtained a score of 86.42% in the classification of conversations of people with dementia and without dementia. This automatic classifier have a better performance than the neural network.

Although we reached a good classification accuracy (according to the nature of the task), we believe that this classification depends a lot on patterns like pauses, daze, confusions and word type. We are not finding patterns related to other mental abilities like understanding of complex content or ability to nominate and categorize. These characteristics are important and should be taken into consideration if we want to improve our analysis.

Acknowledgment

We would like to express our very great appreciation to Dr. Davis Boyd for her valuable support to get the access to the Carolina Corpus Conversations.

References

[1] Curry Guinn and Ben Singer, *A Comparison of Syntax, Semantics, and Pragmatics in Spoken Language among Residents with Alzheimer's Disease in Managed-Care Facilities.*, 2014.

[2] Ali Khodabakshsh and Cenk Demiroglu, *Natural language features for detection of Alzheimer's disease in conversational speech.*, 2014.

[3] B. Roark, M. Mitchell, J. Hosom, K. Hollingshead and J. Kaye, *Spoken Language Derived Measures for Detecting Mild Cognitive Impairment.*, 2011.

[4] Calvin Thomas, Vlado Keselj, Nick Cercone and Kenneth Rockwood, *Automatic Detection and Rating of Dementia of Alzheimer Type through Lexical Analysis of Spontaneous Speech.*, 2005.

[5] Carolinas Conversation Collection, *Carolinas Conversation Collection*, 2019. [Online]. Available:https://carolinaconversations.musc.edu/about/collection. [Accessed: March 11, 2018].

[6] TensorFlow, *TensorFlow*, 2019. [Online]. Available:https://www.tensorflow.org/. [Accessed March 6, 2019].

[7] Keras, *Keras*, 2019. [Online]. Available:https://keras.io/. [Accessed March 6, 2019.]

[8] Scikit-learn, *Scikit-learn*, 2019. [Online]. Available:https://scikit-learn.org/stable/. [Accessed July 15, 2019].

[9] Luis Miguel Gutiérrez-Robledo e Isabel Arrieta-Cruz, *Demencias en México: la necesidad de un Plan de Acción.*, 2015. [Online]. Available:http://www.medigraphic.com/pdfs/gaceta/gm-2015/gm155p.pdf. [Accessed: Feb 15, 2018].

[10] NLTK, *Natural Language Toolkit*, 2019. [Online]. Available:https://www.nltk.org/. [Accessed Feb. 3, 2018].

[11] NLTK, *Natural Language Toolkit*, 2019. [Online]. Available:https://www.nltk.org/. [Accessed Feb. 3, 2018].

[12] Janeth Hernández Jaramillo, *Demencias: los problemas de lenguaje como hallazgos tempranos.*, 2010.

[13] Lai Yi-Hsiu, *Language Processing of Seniors with Alzheimer's Disease: From the Perspective of Temporal Parameters.*, 2017.

[14] S. Bird, E. Klein and E. Loper, *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit.*, 2009.

[15] INEGI, *Población Esperanza de vida.*, 2017. [Online]. Available:http://cuentame.inegi.org.mx/poblacion/esperanza.aspx?tema=P. [Accessed: Feb. 3, 2018].