Automatic Identification of Texts Written by Authors with Alzheimer's Disease

Juan Soler-Company (juan.soler@upf.edu)
NLP Group, Universitat Pompeu Fabra, C/ Roc Boronat 138, 08018 Barcelona, Spain

Leo Wanner (leo.wanner@upf.edu)
NLP Group, Universitat Pompeu Fabra and ICREA, C/ Roc Boronat 138, 08018 Barcelona, Spain

Abstract

As demonstrated in previous studies, Alzheimer's disease leads to a degradation of vocabulary and communication skills. Novels by writers who are known to have suffered from this disease were compared with respect to their lexical richness and syntactic complexity. Those written after the break-out of the disease have shown to use a considerably smaller lexicon and a reduced syntactic complexity of the sentences. This makes us assume that writings of individual authors can be classified automatically into "pre-Alzheimer's period" and "Alzheimer's period". But the writing style of an author is highly individual. Can we still detect whether any given novel is written by an author who suffers from Alzheimer's? To assess this, we use a corpus of novels by three well-known writers who were diagnosed with Alzheimer's: Iris Murdoch, Terry Pratchett and Agatha Christie. Using a mostly stylistic set of features we are able to distinguish between novels written under the influence of the disease and novels written by healthy writers with more than 82% accuracy. The classification of the novels of a given author into "pre-Alzheimer's period" and "Alzheimer's period" is accomplished with more than 86% accuracy. We also prove that our feature set is versatile enough to be able to distinguish between authors in general and books with high precision.

Keywords: Alzheimer's Detection; Text Classification; Author Identification; Author Profiling

Introduction

Alzheimer's disease is a degenerative brain disease and the most common cause of dementia. It is the 6th leading cause of death in the United States and kills more than breast and prostate cancer combined. 1 out of 10 people aged 65 and older suffers from Alzheimer's, so it is quite clear that the impact of the disease in today's society is huge.[1]

The characteristic symptoms of Alzheimer's are difficulties with memory, language, problem solving and other cognitive skills that affect a person's ability to perform everyday activities. People with the disease have trouble following conversations, choosing the right vocabulary and articularing precisely their ideas.

From the natural language processing point of view, the effects of Alzheimer's can be assessed through the analysis of the writing style of an author before and after the break-out of the disease. Several studies in the past carried out such an analysis on novels of well-known authors; cf., e.g., (Garrard et al., 2004; Le et al., 2011; Hirst & Wei Feng, 2012), with the conclusion that a clear decline in vocabulary richness and syntactic complexity and an increase of repetitions after the break-out of the disease can be detected in works during the Alzheimer's period. As a consequence, it can be expected that supervised machine learning techniques will be able to distinguish between the works of an author written before and after the break-out of the disease. However, a more intriguing research question is whether the language patterns of the Alzheimer's disease are generalizable, i.e., whether we can identify if a novel (or text in general) has been written by an author with Alzheimer's or not.

In what follows, we show that indeed the individual works of an author can be classified as belonging to their "pre-Alzheimer's" or "Alzheimer's" period and that a novel can be also identified as being written by an author with Alzheimer's or by an author who does not suffer from the disease. In order to explore further to what extent Alzheimer's leads to a change of style and (possibly also) to a thematic dispersion, we carry out another experiment, in which we automatically assign fragments of different novels to the corresponding novel and author. Furthermore, we analyze the distinctiveness of each feature, to get insight about how the disease affects the style of the authors. Such analysis could be very useful for the implementation of tests that analyze how the writing style of a user changes with time and to warn users when a decline is detected with the goal to detect the disease early and to treat it as effectively as possible.

For our experiments, we retrieved novels from three well-known authors, Iris Murdoch, Agatha Christie and Terry Pratchett, who were extremely productive while being healthy and also wrote some novels under the influence of the disease.

The rest of the paper is structured as follows. The next section reviews the related work. Then, we present the experimental setup, introduce the dataset, the selected features and the results of the implemented experiments. The results are discussed in a separate section. The last section draws some conclusions and outlines our future work.

Related Work

Several works have studied how the Alzheimer's disease affects language. Boyé et al. (2014) study the language of Alzheimer's patients in conversation contexts with known interlocutors. Conversations of five Alzheimer's patients and five control people are analyzed. The conversations are transcribed and lexical, syntactic and spoken features are extracted. The authors study how these features vary depending on whether the subject is a patient, or a control person. The outcome shows that people affected by the disease use fewer words, use more 'yes'/'no' utterances and shorter utterances in general. Paulino and Sierra (2017) looked at interviews conducted with 7 Spanish Alzheimer's disease patients.

[1] https://www.alz.org/facts/

Rhetorical Structure Theory is used to analyze each dialog turn. The results indicate that there are significant differences in the number of rhetoric relations used by Alzheimer's patients when compared to healthy individuals.

Luzzatti et al. (2003) study the results of a writing task given to 23 Italian patients. The study shows that the subjects presented impairment of surface dysgraphia (i.e., the patients cannot access lexical knowledge, but still use phonological-to-orthographic conversion rules correctly, misspelling irregular words), phonological dysgraphia (i.e., patients spell correctly words that they have known how to spell, but cannot spell new words), and in some cases, agraphia (i.e., loss of the ability to write). For further works on the evolution of agraphia and language comprehension; see, e.g., (Cummings & Benson, 1992; Houghton & Zorzi, 2003; Neils-Strunjas, Shuren, Roeltgen, & Brown, 1998).

As already mentioned in the Introduction, some of the studies also analyzed the writings of well-known authors who contracted the disease. See e.g., (Garrard et al., 2004) for an analysis of the works of Iris Murdoch. The authors analyze the syntactic complexity, the lexical variety, the frequency of repetition and the usage of nouns, verbs, descriptors and function words in three novels: her first novel, a novel on her prime, and the novel written under the influence of the disease. Her last novel appears to use simpler syntactic structures and a more restricted vocabulary than the other two studied novels. Le et al. (2011) and (Hirst & Wei Feng, 2012) study lexical and syntactic changes in 26 novels by Iris Murdoch, 16 by Agatha Christie, and 15 by P.D. James (who aged healthily). In this case, several features are studied, namely how vocabulary size, repetition, word specificity, use of passive and use of auxiliary verbs evolve with the disease. The study also shows a clear decline of Iris Murdoch in her last novel, and a more gradual declining tendency in Christie's last novels. (Fraser & Hirst, 2016) analyze the semantic changes in Alzheimer's patients using vector space models. The authors train word representations using healthy control individuals and Alzheimer's patients and analyze the contextual differences of specific words. In conclusion, there are several works that analyze the evolution of linguistic features, but there are none that actually try to automatically distinguish between texts written under the influence of Alzheimer's and texts whose authors do not suffer from Alzheimer's. See also (Chaski, 2012; Koppel, Schler, & Argamon, 2011; O'Brien, 2013) for more generic approaches to authorship attribution using stylometric techniques.

Experimental Setup

In this section, we present the setup of our experiments. We first introduce the corpus on which we carried out the experiments and then the classification features that are used. In the last subsection, we present the experiments and their results.

Dataset

Our corpus is composed of fragments of books by three authors who are assumed to have suffered from Alzheimer's, namely Iris Murdoch, Agatha Christie and Terry Pratchett. For each author, the same number of books written while healthy and under the influence of Alzheimer's have been selected. Each selected book is divided into 300 instances. Depending on the total length of the book, the instances may contain a variable amount of sentences. We ensure that each instance contains full sentences (we do not split sentences between instances).

For Agatha Christie, the selected books are the following: *Curtain*, *Elephants can remember*, and *Sleeping Murder* (written while with Alzheimer's) and *Mysterious Affair at Styles*, *Murder on the Orient Express* and *The Burden* (written while healthy); for Iris Murdoch: *Jackson's Dilemma* (written while with Alzheimer's), and *The Sea* (written while healthy); for Terry Pratchett: *Discworld's 36-37-38-39* (written while with Alzheimer's) and *Discworld 1-2-5-6*(written while healthy). The main reason behind the prominence of Terry Pratchett in our corpus is that he was diagnosed earlier and was able to write more books while suffering from Alzheimer's. Iris Murdoch wrote only one book under the influence of the disease, and even if Agatha Christie has never been officially diagnosed with Alzheimer's, there are clear signs that her last books were much simpler, which has been associated with the neurological decline caused by Alzheimer's; see e.g., (Le et al., 2011; Hirst & Wei Feng, 2012).

Our dataset is thus not completely balanced: 2400 instances are texts by Terry Pratchett, 1800 by Agatha Christie and 600 by Iris Murdoch. However, as we will see later, this does not affect the performance of our classifier.

Feature Set

We implement our experiments as supervised machine learning problems in which a set of features is extracted to characterize an instance with respect to its label. We use Weka's implementation of LibSVM (Hall et al., 2009) with a linear kernel for classification and 10-fold cross validation in order not to be biased by the selection of a training respectively test data subset. The feature set is composed of six subgroups of features introduced below; for their extraction, we use Python's natural language toolkit and Bohnet and Nivre (2012)'s dependency parser. Raw text is converted into multidimensional vectors, where each dimension is a feature.

The feature set is composed of six subgroups of features introduced below.

Character-based Features are composed of the ratios between upper cased characters, periods, commas, parentheses, exclamations, colons, number digits, semicolons, hyphens and quotation marks and the total number of characters in a text.

Word-based Features are composed of the mean values of characters per word, vocabulary richness, acronyms, stopwords, first person pronouns, usage of words composed by two or three characters, standard deviation of word length and the difference between the longest and shortest words.

Sentence-based Features are composed of the mean number of words per sentence, standard deviation of words per sentence and the difference between the maximum and minimum number of words per sentence in a text.

Dictionary-based Features consist of the ratios of discourse markers, interjections, abbreviations, curse words, polar words (positive and negative words using the polarity dictionaries described in (Hu & Liu, 2004)) and emotion words with respect to the total number of words in a text. The emotion word features are computed using a publicly available resource called "Depeche Mood", which provides dictionaries that contain words that evoke the following emotions: fear, amusement, anger, annoyance, indifference, happiness, inspiration and sadness; for more information, refer to (Staiano & Guerini, 2014). For each one of these emotions, two features are computed: the mean number of words per text that correspond to each specific emotion and the percentage of the emotion words that belong to that particular emotion. The mean ratio of emotion words per text in general is also computed.

Syntactic Features Three types of syntactic features are distinguished:

1. *Part-of-Speech Features* are given by the relative frequency of each PoS tag[2] in a text, the relative frequency of comparative/superlative adjectives and adverbs and the relative frequency of the present and past tenses. In addition to the fine-grained Penn Treebank tags, we introduce general grammatical categories (such as 'verb', 'noun', etc.) and calculate their frequencies.

2. *Dependency Features* reflect the occurrence of syntactic dependency relations in the dependency trees of the text. The dependency tagset used by the parser is described in (Surdeanu, Johansson, Meyers, Màrquez, & Nivre, 2008). We extract the frequency of each individual dependency relation per sentence, the percentage of modifier relations used per tree, the frequency of adverbial dependencies (they give information on manner, direction, purpose, etc.), the ratio of modal verbs with respect to the total number of verbs, and the percentage of verbs that appear in complex tenses referred to as "verb chains" (VCs).

3. *Tree Features* measure the tree width, the tree depth and the ramification factor of the tree. Tree depth is defined as the maximum number of nodes between the root and a leaf node, the width is the maximum number of siblings at any of levels of the tree, and the ramification factor is the mean number of children per level. In other words, the tree features characterize the complexity of the inner structure of the sentences. These measures are also applied to subordinate and coordinate clauses.

Analyzing how these metrics evolve with respect to the health status of an author can give us an idea on whether the complexity of the syntactic structures decreases as the disease progresses or not.

Lexical Features (or content-dependent features) are used to complement our mainly structural/stylistic features. This group contains the frequencies of the 50 most frequent words of our corpus.

Our full set of features consists thus of less than 200 features, which, compared with most of the state-of-the-art works on author identification/profiling and on text classification in general, is rather low (and still obtains state-of-the-art performance). Earlier versions of the feature set have been successfully used in several tasks (see e.g., (Soler-Company & Wanner, 2017b, 2015, 2017a)), and we believe that the current version is general enough to tackle different tasks effectively, so it is an appropriate fit for the problem at hand.

To contrast the performance of our feature set, two baselines are chosen. The first one is very simple, the majority class baseline, which classifies every instance as the class with more instances in the corpus, showing how challenging an experiment really is. The second one is a token bigram (sequences of two consecutive words) baseline, which uses the frequencies of the most frequent 100, 300, 500, 700 and 900 bigrams for classification. We also considered using trigrams and 4-grams, but their performance was worse than that of bigrams in all cases, so they were discarded.

Experiments and Results

We carried out several experiments. The first batch of experiments aims to identify, given a text instance, the author, the book and whether the author of the text has Alzheimer's or not. In these experiments, the full dataset is used. The second batch of experiments tries to distinguish between each author when healthy vs. the same author when ill. In each experiment from the second batch, only instances of the specific author are used. For each experiment, we present the performance of our full set of features, of each feature group by itself and of both baselines.

The results of the first batch of experiments are shown in Table 1.

Table 1: Results of the first set of experiments.

Features Used	Author Id	Alzheimer's Id	Book Id
Full Set	**96,39%**	**82,21%**	**73,02%**
Character-based	69,75%	64,91%	35,04%
Word-based	83,44%	70,60%	42,65%
Sentence-based	61,65%	60,85%	19,25%
Dictionary-based	71,58%	65,56%	33,33%
Syntactic	94,71%	73,83%	55,15%
Lexical	57,45%	54,47%	19,98%
Majority Class	50%	50%	6%
Token 2-gram 100	80,23%	68,44%	33,27%
Token 2-gram 300	84,15%	72,52%	40,39%
Token 2-gram 500	85,87%	74,60%	48,06%
Token 2-gram 700	89,47%	76,66%	52,94%
Token 2-gram 900	90,87%	78,01%	57,35%

[2]We use the Penn Treebank tagset http://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

The results of the second batch of experiments are presented in Table 2.

Table 2: Results of the second set of experiments.

Features Used	Iris Murdoch	Agatha Christie	Terry Pratchett
Full Set	98,50%	86,00%	94,50%
Character-based	82,33%	72,67%	76,63%
Word-based	87,17%	68,51%	85,00%
Sentence-based	68,67%	59,33%	69,29%
Dictionary-based	74,50%	66,50%	76,50%
Syntactic	89,51%	76,89%	86,21%
Lexical	75,83%	71,61%	67,13%
Majority Class	50%	50%	50%
Token 2-gram 100	85,67%	64,50%	83,71%
Token 2-gram 300	87,01%	65,17%	85,63%
Token 2-gram 500	88,55%	65,94%	85,33%
Token 2-gram 700	90,19%	65,39%	87,95%
Token 2-gram 900	90,66%	69,17%	88,78%

Discussion

Table 1 shows the performance in the first batch of experiments. It can be observed that the performance of our full set of features is competitive, achieving more than 96% of accuracy in author identification, more than 82% in Alzheimer's identification and finally, in the most challenging experiment, the book identification case (where the majority class baseline is only 6%), 73,02%. In each case, the classifier with our features is able to outperform the baselines. The table also shows the performance of each individual set of features in each experiment. Some conclusions can be drawn from the performance of the individual feature groups. In all cases, the syntactic group of features performs best; it is also the largest group, and the one that best characterizes the writing style of the authors, without analyzing specific choices of words. We see that the baseline achieves good performances in author and Alzheimer's identification, but has a harder time in the book classification case. It needs to be noted that the best performances of the baseline involve the use of 900 features, which is a much larger number of features compared to our feature set. We can also observe that the lexical features are not very effective by themselves and that word-based features obtain competitive performance in author and Alzheimer's identification, which can be due to the fact that this group of features analyzes the characteristics of words and the vocabulary richness of the authors, one of the characteristics that can directly be related to the cognitive degradation that the disease causes.

Figure 1 shows the confusion matrix of the book identification experiment. In general, this matrix shows that the feature set captures effectively the style of an author and that books by the same author are often confused between each other, while books by different authors are confused very infrequently. More specifically, the matrix indicates that Discworld 36 and 37 and Discworld 38 and 39 are often confused between each other. It is also notable that even though Discworld 37 is often confused (in particular) with 36, 38-39 and 6 (not as frequently), it is never confused with Discworld 1 and 2, and only once with Discworld 5. This shows a clear evolution of the writing style of Terry Pratchett during the development of the saga. It also shows that books written under the influence of Alzheimer's are stylistically similar enough to be confused with each other often. The case of Iris Murdoch shows that the book written with Alzheimer's and the one written while healthy are confused only in two cases, which shows how different stylistically these two books are. The books by Agatha Christie are mostly confused between each other, which shows the consistency of the style of the author even with the disease. In other words, the style of some authors tends to change significantly or become less distinctive from work to work during the period they suffered from Alzheimer, while the style of others remained stable.

Table 2 shows the performance of the second batch of experiments. This experiment aimed to distinguish between the writings of the same author when healthy and when ill. The table shows that the performance of our classification is rather competitive in this case as well, with more than 86% of accuracy in all cases. In the Iris Murdoch case, we obtain an accuracy of 98,50%, which is almost perfect. This can be due to the fact that there are only 300 instances per class in this case and the two selected books are very different stylistically. However, looking at the performance of the Terry Pratchett experiment, we can see that we obtain 94,50% of accuracy while distinguishing books from the same author and saga, sharing themes, characters, and universe, which makes the classification task much more challenging. In all cases, we outperform the baselines by a large margin.

One of the main advantages that our (mainly) stylistic features have against other feature sets such as word embeddings, bag-of-words approaches or other content-based features, is that we can analyze the values of many different linguistic features in different settings. This analysis can provide very valuable information on the effects of the disease on the writing style of the analyzed authors. Computing the information gain of the features in each one of the Alzheimer's-related experiments, we can see the features that were the most relevant for the classification. Table 3 shows the 10 most distinctive features in each of the Alzheimer's-related experiments. Features with 'SYNPOS' as prefix represent part-of-speech frequencies, the ones with 'SYNDEP' are dependency relation frequencies and 'SYNSHAPE' are shape-based metrics of the dependency trees

For convenience of the reader, we list the definitions of these features:[3]

- SYNPOS_POS: Word with possessive ending,
- SYNDEP_PRT: Particle (dependent on verb),
- SYNPOS_WP: Wh-pronoun,
- SYNPOS_RP: Particle,

[3]For the definition of the full list of dependencies, see (Surdeanu et al., 2008).

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	<-- classified as
210	27	0	1	14	0	0	2	0	0	0	3	18	25	0	0	a = sleepingmurder
33	209	0	5	16	0	2	1	3	2	1	8	6	11	1	2	b = theBurden
0	0	285	0	0	0	0	2	1	0	1	0	0	0	4	7	c = theSeatheSea
1	7	0	212	4	38	24	7	1	3	0	0	0	0	0	3	d = discworld1
15	34	1	3	203	0	0	0	1	1	0	13	14	15	0	0	e = murderorientexpress
0	1	0	27	0	163	63	20	1	5	9	1	0	1	5	4	f = discworld2
1	1	2	23	0	59	169	26	0	2	7	0	0	1	2	7	g = discworld5
0	2	2	4	0	12	27	218	2	0	5	0	0	0	13	15	h = discworld6
1	2	1	1	0	0	1	5	268	4	6	0	0	6	2	3	i = jacksonDilemma
0	0	0	2	0	0	9	0	1	219	43	1	1	0	11	13	j = discworld39
1	0	1	1	0	0	0	10	2	41	228	0	0	0	5	11	k = discworld38
4	19	0	2	30	1	0	0	0	0	0	217	8	19	0	0	l = misteriousaffairatstyles
20	16	0	1	17	0	0	0	0	1	1	6	236	2	0	0	m = elephantsCanRemember
22	17	0	4	10	0	0	0	8	0	0	10	5	224	0	0	n = curtain
0	0	7	0	0	0	1	9	0	8	6	0	0	0	226	43	o = discworld37
0	2	2	2	0	0	3	11	2	12	10	0	0	0	38	218	p = discworld36

Figure 1: Confusion matrix of the book identification experiment.

Table 3: 10 features with more information gain in every Alzheimer's-related experiment.

Alzheimer's Id	Iris Murdoch	Agatha Christie	Terry Pratchett
SYNPOS_POS	SYNDEP_compVerbRatio	SYNPOS_VBP	SYNDEP_OPRD
SYNDEP_PRT	SYNDEP_MNR	SYNDEP_OPRD	SYNDEP_IM
SYNPOS_WP	SYNDEP_APPO	SYNDEP_IM	SYNDEP_SUB
SYNPOS_RP	SYNPOS_RP	SYNDEP_compVerbRatio	SYNPOS_PRP$
SYNDEP_OPRD	SYNDEP_PRT	SYNDEP_MNR	SYNPOS_MD
SYNPOS_MD	SYNDEP_DIR	SYNPOS_MD	SYNPOS_POS
SYNPOS_VBP	SYNPOS_WRB	SYNPOS_VBD	SYNDEP_APPO
SYNDEP_SUB	SYNPOS_WP	SYNDEP_LOC	SYNPOS_VBP
SYNDEP_MNR	SYNPOS_POS	SYNPOS_VBG	SYNDEP_MNR
SYNPOS_WRB	SYNDEP_LOC	SYNDEP_AMOD	SYNPOS_WP

- SYNDEP_OPRD: Predicative complement of raising/control verb,
- SYNPOS_MD: Modal verb,
- SYNPOS_VBP: Verb, non-3rd person singular present,
- SYNDEP_SUB: Subordinated clause,
- SYNDEP_MNR: Adverbial of manner,
- SYNPOS_WRB: Wh-adverb,
- SYNDEP_compVerbRatio: ratio of composed verbs vs. total number of verbs,
- SYNDEP_APPO: Apposition,
- SYNDEP_DIR: Adverbial of direction,
- SYNDEP_LOC: Locative adverbial,
- SYNDEP_IM: Infinitive verb (dependent on infinitive marker to),
- SYNPOS_VBD: Verb, past tense,
- SYNPOS_VBG: Verb, gerund or present participle
- SYNDEP_AMOD: Modifier of adjective or adverbial,
- SYNPOS_PRP$: Possessive pronoun.

Note that the features that are displayed in Table 3 show the most distinctive features in each experiment considering the full set of features. As we see, all of these features are syntactic, showing that the analysis of the syntactic traits is a good way to measure the stylistic evolution of an author. The first non-syntactic feature that appears in this feature ranking is the vocabulary richness, which is also a good indicator of the lexical variety that an author shows throughout different moments of his/her career. If we analyze the specific syntactic features that are distinctive, we see that for the general case (Alzheimer's Id) and for the case of Terry Pratchett, the number of subordinate clauses is very distinctive. This could mean that complex structures such as subordinate clauses are found more scarcely in the texts written by authors with Alzheimer's. Other features such as the ratio of composed verbs and the usage of adverbial dependencies (which indicate manner, location, direction, etc.) are also very distinctive. The ratio of composed verbs and the usage of adverbial dependencies are features that indicate that a text gives detailed, precise explanations (specifying locations, manners, directions, purpose, or extent) and uses complex verb structures. A decline of these features could indicate a decline of the writing style of the author.

Conclusions and Future Work

This paper presents classification experiments on the distinction between the writings of authors with Alzheimer's and

healthy authors. We show that it is possible to differentiate between the writings of the same author with and without the disease very effectively, and even more: that it is possible to identify whether a novel has been written by an author who suffers from Alzheimer's or by an author who does not. Our book identification experiments, in which we assigned isolated text instances to specific novels, have shown that the style of a writer may change with Alzheimer's and become less distinctive from work to work or remain stable. Further, broader studies are needed to investigate this issue in more depth.

From the perspective of feature engineering, we analyze the features that are most distinctive in all Alzheimer's disease classification experiments, showing the relevance of syntactic features in the experiments and relating them to the development of the disease. We also analyze the confusions that emerge from the book identification experiment, which prove that with the chosen features we are effectively capturing the writing style of the authors.

In the future, we plan to expand this work using data from patients to see whether these stylistic patterns also appear in non-literary texts. We also plan to explore different feature sets and approaches, using texts written in different languages.

References

Bohnet, B., & Nivre, J. (2012). A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 1455–1465).

Boyé, M., Tran, T. M., & Grabar, N. (2014). Nlp-oriented contrastive study of linguistic productions of alzheimers and control people. In *International conference on natural language processing* (pp. 412–424).

Chaski, C. E. (2012). Best practices and admissibility of forensic author identification. *JL & Pol'y*, *21*, 333.

Cummings, J. L., & Benson, D. F. (1992). *Dementia: A clinical approach*. Butterworth-Heinemann Medical.

Fraser, K. C., & Hirst, G. (2016). Detecting semantic changes in alzheimers disease with vector space models. In *Proceedings of lrec 2016 workshop. resources and processing of linguistic and extra-linguistic data from people with various forms of cognitive/psychiatric impairments (rapid-2016)*.

Garrard, P., Maloney, L. M., Hodges, J. R., & Patterson, K. (2004). The effects of very early alzheimer's disease on the characteristics of writing by a renowned author. *Brain*, *128*(2), 250–260.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, *11*(1), 10–18.

Hirst, G., & Wei Feng, V. (2012). Changes in style in authors with alzheimer's disease. *English Studies*, *93*(3), 357–370.

Houghton, G., & Zorzi, M. (2003). Normal and impaired spelling in a connectionist dual-route architecture. *Cognitive neuropsychology*, *20*(2), 115–162.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth acm sigkdd international conference on knowledge discovery and data mining* (pp. 168–177). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/1014052.1014073` doi: 10.1145/1014052.1014073

Koppel, M., Schler, J., & Argamon, S. (2011). Authorship attribution in the wild. *Language Resources and Evaluation*, *45*(1), 83–94.

Le, X., Lancashire, I., Hirst, G., & Jokel, R. (2011). Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three british novelists. *Literary and Linguistic Computing*, *26*(4), 435–461.

Luzzatti, C., Laiacona, M., & Agazzi, D. (2003). Multiple patterns of writing disorders in dementia of the alzheimer type and their evolution. *Neuropsychologia*, *41*(7), 759–772.

Neils-Strunjas, J., Shuren, J., Roeltgen, D., & Brown, C. (1998). Perseverative writing errors in a patient with alzheimer's disease. *Brain and Language*, *63*(3), 303–320.

O'Brien, S. (2013). The borrowers: Researching the cognitive aspects of translation. *Target. International Journal of Translation Studies*, *25*(1), 5–17.

Paulino, A., & Sierra, G. (2017). Applying the rhetorical structure theory in alzheimer patients' speech. In *Proceedings of the 6th workshop on recent advances in rst and related formalisms* (pp. 34–38). Association for Computational Linguistics.

Soler-Company, J., & Wanner, L. (2015). Multiple language gender identification for blog posts. In *Proceedings of the 37th annual meeting of the cognitive science society* (pp. 2248–2253).

Soler-Company, J., & Wanner, L. (2017a). On the relevance of syntactic and discourse features for author profiling and identification. In *European chapter of the association for computational linguistics, eacl 2017* (pp. 681–687).

Soler-Company, J., & Wanner, L. (2017b). On the role of syntactic dependencies and discourse relations for author and gender identification. *Pattern Recognition Letters*.

Staiano, J., & Guerini, M. (2014). Depechemood: a lexicon for emotion analysis from crowd-annotated news. *CoRR*, *abs/1405.1605*.

Surdeanu, M., Johansson, R., Meyers, A., Màrquez, L., & Nivre, J. (2008). The conll-2008 shared task on joing parsing of syntactic and semantic dependencies. In *Proceedings of the twelfth conference on computational natural language learning* (pp. 159–177). Stroudsburg, PA, USA: Association for Computational Linguistics.