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

**Juan Soler-Company (juan.soler@upf.edu)**
Universitat Pompeu Fabra, Carrer de Roc Boronat 138
Barcelona, 08018, Spain

**Leo Wanner (leo.wanner@upf.edu)**
Universitat Pompeu Fabra and ICREA, Carrer de Roc Boronat 138
Barcelona, 08018, Spain

## Abstract

Alzheimer's disease is the most common cause of dementia. One of the characteristic symptoms of the disease is the degradation of vocabulary and communication skills. To analyze this symptom, we used a corpus of well-known novelists who were diagnosed with Alzheimer's and used texts written before and after contracting the disease. The three chosen authors are Iris Murdoch, Terry Pratchett and Agatha Christie. Using a mostly stylistic set of features we were able to distinguish between texts written under the influence of the disease and texts written while healthy with more than 82% accuracy. We also were able to classify texts of each author depending on whether the text was written under the influence of the disease or not, obtaining more than 86% accuracy in every case. Our approach is also able to distinguish between books and between authors with high accuracy. Finally a feature analysis process is shown, where we analyze which features are the most distinctive in each Alzheimer's classification experiment.

**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 has Alzheimer's disease, so it is quite clear that the impact of the disease in today's society is huge[1].

The characteristic symptoms of the disease are difficulties with memory, language, problem solving and other cognitive skills that affect a persons ability to perform everyday activities among many others. People with Alzheimer's disease have trouble following conversations, struggle with vocabulary and have problems expressing themselves with precision.

From the natural language processing point of view, an appropriate way to study the effects of the disease is to analyze the writing style of an author before and after the author contracts it. To perform such a study, we retrieved novels from well-known authors, which wrote many books while being healthy, and also wrote some of their novels under the influence of the disease. We chose three authors: Iris Murdoch, Agatha Christie and Terry Pratchett and aimed to automatically distinguish between the writings of the authors when healthy and when ill. Each one of the authors has a very distinguishable writing style and a decline in their neurological capacities can be translated into a noticeable variation of their writing style, which can be automatically detected using stylistic analysis techniques and machine learning.

In this paper, we present an automatic approach that is able to distinguish between the writings of the chosen authors, between books, and between the writings of each author when healthy and when ill. Furthermore, we analyze the distinctiveness of each feature to get insight on how the disease affects the style of the authors. Such analysis could be very useful to implement tests that analyze how the writing style of a user changes with time and to warn users when a decline is detected. This application could be an initial warning to detect the disease early and to treat it as effectively as possible.

The rest of the paper is structured as follows: Section  reviews the related work, Section  presents the experimental setup, introduces the dataset, the selected features and the results of the implemented experiments, Section  discusses the results, and finally Section  draws some conclusions and outlines our future work.

## Related Work

Several works have studied how Alzheimer's disease affects language. (Boyé, Tran, & Grabar, 2014) study the language of Alzheimer's patients in conversation contexts with known interlocutors. Conversations of five Alzheimer's disease patients and five control people are studied. 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, showing that people affected by the disease use fewer words, use more yes/no utterances and shorter utterances in general. The authors of (Paulino & Sierra, 2017) analyze interviews conducted with 7 Spanish Alzheimer's disease patients. 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 disease patients with respect to healthy individuals.

(Luzzatti, Laiacona, & Agazzi, 2003) study the results of a writing task given to 23 Italian patients suffering from Alzheimer's disease. The study showed that the subjects presented impairment of surface dysgraphia (patients cannot access lexical knowledge but still use phonological-to-orthographic conversion rules correctly, misspelling irregular words), phonological dysgraphia (patients spell correctly words that they have known how to spell, but cannot spell

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

new words), and in some cases, agraphia (loss in the ability to write). Further works focused on the evolution of agraphia and language comprehension can be found in (Cummings & Benson, 1992; Houghton & Zorzi, 2003; Neils-Strunjas, Shuren, Roeltgen, & Brown, 1998).

Previous works also analyze the writings of well-known authors that contracted the disease. See e.g., (Garrard, Maloney, Hodges, & Patterson, 2004) for an analysis of the works of Iris Murdoch. The authors analyze the syntactic complexity, the lexical variety, how frequently repetition occurs and the usage of nouns, verbs, descriptors and function words, using 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 than the other studied novels and to use a more restricted vocabulary.

Similar approaches can be found in (Le, Lancashire, Hirst, & Jokel, 2011; Hirst & Wei Feng, 2012), where the authors study lexical and syntactic changes using 26 novels from 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. In a more recent approach, (Fraser & Hirst, 2016) (some) of the same authors analyze the semantic changes on Alzheimer's disease patients using vector space models. The authors train word representations using healthy control people and Alzheimer's patients and analyze the contextual differences on specific words.

As we can see, there are several works that analyze the evolution of linguistic features, but there are not many approaches that actually try to automatically classify between texts written under the influence of the disease or while healthy.

## Experimental Setup

In this section we present the experimental setup. Firstly we present the corpus that is used in every experiment. Secondly we describe the feature set that is used in every classification experiment. Finally, we present the experiments and their results.

### Dataset

The corpus that is used was compiled for this paper and is composed of 4800 texts. These texts correspond to writings of three authors that have been diagnosed with Alzheimer's (or at least, there is a general consensus that they conceived the disease), namely Iris Murdoch, Agatha Christie and Terry Pratchett. For each author, the same amount of books written while healthy and under the influence of Alzheimer's disease have been selected. Each selected book is divided in 300 instances, which contain a variable amount of sentences (depending on the book's length). We ensure that each instance contains full sentences (we do not split sentences between instances). Given the nature of the texts that compose the corpus, we believe that the stylistic changes will be clearer and much more illustrative, than using for example, e-mails written by Alzheimer's disease patients before and after contracting the disease (and of course, such texts are much more difficult to obtain).

For Agatha Christie, the selected books are the following: Curtain, Elephants can Remember and Sleeping Murder (with Alzheimer's) and Mysterious Affair at Styles, Murder on the Orient Express and The Burden (healthy). For Iris Murdoch, Jackson's Dilemma (with Alzheimer's), and The Sea, the Sea (healthy). Finally, for Terry Pratchett, Discworld's 36-37-38-39 are selected as books written with Alzheimer's and Discworld 1-2-5-6 as healthy-written books. As we can see, the dataset is not completely balanced, 2400 instances are texts by Terry Pratchett, 1800 by Agatha Christie and 600 by Iris Murdoch. As we will see later, even though the dataset is skewed, our system performs very competitively. 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 sick. Iris Murdoch only wrote a book under the influence of the disease. Finally, even though Agatha Christie has never been officially diagnosed, there are clear signs that her last books were much simpler and that has been associated with the neurological decline caused by Alzheimer's disease (see e.g., (Le et al., 2011; Hirst & Wei Feng, 2012)).

### Feature Set

The experiments we are going to present are supervised machine learning problems in which a feature set is extracted to characterize an instance with respect to its label. We use Support Vector Machines (Weka's implementation of LibSVM) with a linear kernel for classification and 10-fold cross validation for the results to not depend on the chosen subsets of training and test set.

For feature extraction, Python, its natural language toolkit and a dependency parser (Bohnet & Nivre, 2012) are used. 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**

This group of features captures the usage of punctuation marks: commas, periods, parenthesis, exclamation and question marks, hyphens, colons, semi-colons, quotations and other symbols such as the percentage sign, the ampersand and the plus sign. These features are calculated as the ratio of the number of apparitions of said characters in each instance and the total number of characters. The percentage of upper case characters and the usage of numbers is also computed.

**Word-based features**

Word-based features are composed by the mean number of characters per word, the standard deviation in word length and the difference between the longest and shortest word, the vocabulary richness and the usage of acronyms, stop words and first person pronouns, both singular and plural (as ratios of these values and the total number of words in the text).

**Sentence-based features**

In this group of features, three features are computed: the mean number of words per sentence, the standard deviation in sentence length and the difference between the longest and shortest sentence.

**Dictionary-based features**

Dictionary-based features use several different dictionaries: interjections, discourse markers, positive/negative words, abbreviations, curse words and emotion words. The ratios of found words for each dictionary and the total number of words of the text are used as features. The positive and negative word dictionaries are sentiment analysis lexicons used in (Hu & Liu, 2004), which are publicly available. The emotion word lexicons belong to 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. We compiled the other mentioned lexicons.

**Syntactic Features**

This group of features accounts for more than 65% of the total number of features. We can subdivide this group of features into three subgroups:

*Part of Speech*

This subgroup of features contains the frequency of each parts-of-speech tag. We compute the percentage of words of a text that are classified as each one of the possible morphosyntactic categories.[2]

These features can be very useful for the analysis of the distribution of word categories per text. A higher usage of adjectives could be seen as an indicator on the expressiveness of a text. The analysis of a text based on this kind of features can help us find patterns that are author-specific and that can help distinguishing between texts written by different authors or different health statuses of the same author.

*Dependency Features*

This subgroup of features uses the output of the dependency parser mentioned before. Each sentence is represented by a dependency tree where the arcs are syntactic relations between words. The dependency tags delivered by the parser are described in (Surdeanu, Johansson, Meyers, Màrquez, & Nivre, 2008).

From the dependency trees, we extract the frequency of each one of the individual dependency relations per sentence, the percentage of modifier relations used per tree as well as 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 compose a verb chain (such as "has taken", "were thinking", etc) are also part of this feature group.

*Tree-Shape Features*

The goal of this subgroup of features is to extract information from the shape of the dependency trees. We measure their width, depth and ramification factor. The depth is the maximum number of nodes between the root and a leaf node. The width is the maximum number of siblings at a level of the tree. The ramification factor is the mean number of children per level. These features characterize the complexity of the inner structure of the sentences.

We also apply these measures to subordinate and coordinate clauses, whose existence indicates that a sentence has a certain degree of complexity. When we complement the complexity figures of the clauses with the figures concerning their shape, we measure exactly how complex these subtrees are.

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**

To complement our mainly structural/stylistic features, we also use some content-dependent features, namely, lexical features. This group contains the frequencies of the 50 most frequent words of our corpus.

The full set of features consists in less than 200 features, which compared with some of the works in the state-of-the-art of author identification/profiling and in general, in text classification, is rather low. Earlier versions of the feature set have been successfully used in several tasks (Soler-Company & Wanner, 2014, 2015, 2017) and we believe that the current version is general enough to tackle different tasks effectively, so it is a very good 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 baseline (sequences of two consecutive words), which uses the frequencies of the most frequent 100, 300, 500, 700 and 900 bigrams to classify. We also considered using trigrams and 4-grams, but their performance was worse in every case than bigrams, so they were discarded.

## Experiments and Results

Several experiments were implemented. The first set of experiments aims to identify, given a text, the author, the book and if the author of the text has Alzheimer's disease or not. In these experiments, the full dataset is used. The second set of experiments tries to distinguish between each author, when healthy vs. the same author when ill. In each one of these experiments, only instances of the specific author are used. For every experiment, we present the performance of our full set of features, of each feature group by itself and of both baselines.

[2]The tag set that was used can be found in *http://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html*

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

Table 1: Results of the first set of experiments.

| Features Used | Author Id | Alzheimer Id | Book Id |
|---|---|---|---|
| Full Set | **96,39%** | **82,21%** | **73,02%** |
| Char | 69,75% | 64,91% | 35,04% |
| Word | 83,44% | 70,60% | 42,65% |
| Sent | 61,65% | 60,85% | 19,25% |
| Dict | 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% |

The results of the second set 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% |
| Char | 82,33% | 72,67% | 76,63% |
| Word | 87,17% | 68,51% | 85,00% |
| Sent | 68,67% | 59,33% | 69,29% |
| Dict | 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 set of experiments. We can see how the performance of our full set of features is quite competitive, achieving more than 96% of accuracy in author identification, more than 82% in Alzheimer's disease identification and finally, in the most challenging experiment, the book identification case (where the majority class baseline is only 6%), 73,02%. In every case, our selected feature set is able to outperform the baselines. The table also shows the performance of each individual set of features in every experiment. Some conclusions can be drawn from the performance of the feature groups by themselves: in every case, 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 can 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 using 900 features which is a much higher number than our feature set. We can also see how 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.

In the book classification experiment, analyzing the mistakes the classifier makes provides useful insight on the stylistic characterization capability of our system. To do this analysis, we can look at the confusion matrix to see the book misclassifications. The following image shows the confusion matrix of the book identification experiment.

From the figure, we can see how Discworld 36 and 37 and Discworld 38 and 39 are often confused with one another. It is also notable that even though Discworld 37 is often confused with 36 (specially), 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 and how books written under the influence of Alzheimer's are stylistically similar enough to be confused with one another often. The case of Iris Murdoch shows how the book written with Alzheimer's and the one written while healthy are only confused 2 times, which shows how different stylistically these two books are. The books by Agatha Christie are mostly confused with one another, which shows the consistency of the style of the author even with the disease. The matrix, in general, shows how the feature set captures effectively the style of an author, showing how books by the same author are often confused with one another while books by different authors are confused very infrequently.

Table 2 shows the performance of the second set of experiments. This experiment aimed to distinguish between the writings of the same author when healthy and when ill. The table shows how our approach performs very competitively, achieving more than 86% of accuracy in every case. 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 how our system obtains 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 every case, our approach outperforms the baselines by a large margin.

One of the main advantages that our (mainly) stylistic feature set has against other feature sets such as word embeddings, bag-of-words approaches or other content-based approaches, is that we can analyze the values of many different linguistic features in different settings. The analysis of the

| 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.

distinctiveness of our feature set can give 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 in the classification process. Table 3 shows the 10 most distinctive features in each one 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

Even though the full definitions of Part-of-speech tags and dependencies can be found at (Surdeanu et al., 2008) and `http://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html`, we provide the meanings of the most informative features for completion:

- SYNPOS_POS: Word with possessive ending,
- SYNDEP_PRT: Particle (dependent on verb),
- SYNPOS_WP: Wh-pronoun,
- SYNPOS_RP: Particle,
- 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, every one of these features is from the syntactic group, 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 on 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 can see that for the general case (Alzheimer 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 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 explanations (specifying locations, manners, directions, purpose, or extent) and that uses complex verb structures. A decline on these features could indicate a decline on the writing style of the author.

## Conclusions and Future Work

This paper presents an approach to distinguish between the writings of authors with Alzheimer's and healthy authors. We show how our system is able to differentiate between the writings of the same author with and without the disease very effectively. We also present author and book identification experiments.

We analyze the features that are the most distinctive in each Alzheimer's disease classification experiment, showing the relevance of syntactic features in the experiments and relating them with the development of the disease. We also analyzed

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

| Alzheimer 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 |

the confusions that emerged from the book identification experiment, which proved that our system was effectively capturing the writing style of the authors.

In the future work we want to expand this work using data from patients to try and see if these stylistic patterns also appear in non-literary texts. We also want 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).

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.

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

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.

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. (2014). How to use less features and reach better performance in author gender identification. In *The 9th edition of the language resources and evaluation conference (lrec)* (pp. 26–31).

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. (2017). 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).

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.