Computational Linguistics in the Netherlands Journal 8 (2018) 66–82 Submitted 06/2018; Published 12/2018

The Claus Case: Exploring the Use of Propositional Idea Density for Alzheimer Detection

Silke Marckx SILKE-MARCKX@HOTMAIL.COM
Ben Verhoeven BEN.VERHOEVEN@UANTWERPEN.BE
Walter Daelemans WALTER.DAELEMANS@UANTWERPEN.BE

CLiPS Research Center, University of Antwerp, Prinsstraat 13, Antwerp, Belgium

Abstract

Low linguistic ability has been associated with low cognitive reserve, which might result in the development of Alzheimer's disease. As a result, propositional idea density (PID), as a measure of linguistic ability in early life, might predict cognitive decline in late life. This paper explores the differences in propositional behavior between healthy individuals and Alzheimer's patients, using a newly developed computerized propositional idea density measure for Dutch texts. This exploratory study describes an experiment on literary text. We measured the propositional idea density of the works of one author without Alzheimer's disease (i.e. Elsschot) and one author with attested Alzheimer's disease (i.e. Claus). Changes in propositional idea density for both authors were compared, as well as the differences in propositional idea density in early life. Analyses from this experiment showed that the propositional idea density in early life of Elsschot was not significantly higher than that of Claus. The propositional idea density of Elsschot significantly increased over time. This change in propositional idea density greatly differed from the slight decrease in propositional idea density of Claus. On the one hand, this study fails to support the hypothesis that a low propositional idea density in early life predicts cognitive decline in late life. On the other hand, the results provide support to the hypothesis that a slight decrease in propositional idea density over time might be a predictor of cognitive decline in late life. However, much more research is needed to corroborate these findings. The propositional idea density software for Dutch is available on request.

1. Introduction

Language behavior is one of the first cognitive functions that is affected by Alzheimer's disease (AD) as a result of the early development of cognitive dysfunction (Hamilton 1989). Overall, semantic and pragmatic language abilities appear to be more affected by the disease, whereas the syntactic and phonemic features are relatively spared (Croisile et al. 1996). Longitudinal studies (Elias et al. 2000, Kluger et al. 1999) pointed out that, before the onset of Alzheimer's disease, patients already show disturbed cognitive functions that remain relatively stable for several years before the diagnosis of AD is established. The stage in which the patient suffers from cognitive impairments reveals significant similarities with the clinical concept of Mild Cognitive Impairment (MCI). Petersen et al. (2001) describe MCI as "the state of cognition and functional ability between normal aging and very mild AD". People are diagnosed with MCI when one or more cognitive functions are disturbed, but without meeting the clinical criteria for dementia or Alzheimer's disease. However, MCI subjects are far more likely to develop Alzheimer's disease in the future (Petersen et al. 2001). It is important to keep in mind that, besides Alzheimer's patients, healthily aging adults may suffer cognitive decline as well. However, there is a significant difference between the language decline of normal aging adults and Alzheimer's patients. The main clinical distinction between language decline as a result of normal aging and a disease-related language decline is the rate of change. Language decline is much more gradual and less severe in the former, whereas the rate of change in the latter is much more rapid and progressive (Le et al. 2011, Tarawneh and Holtzman 2012).

©2018 Silke Marckx, Ben Verhoeven and Walter Daelemans.

For the last decades, there has been a great interest in the changes in propositional idea density, as a measure of language ability, caused by AD. The concept of propositional idea density dates from more than four decades ago. Kintsch et al. (1973, 1978) first introduced the idea that the content of a text is stored in memory in terms of underlying semantic structures. These abstract structures are represented as propositions. Later on, the method of propositional analysis of texts was operationalized by Turner and Greene (1977) in the form of a practical manual. Propositions are idea units consisting of word concepts that serve as a relation, and the arguments of the proposition. Each proposition reflects one idea, but can be embedded as the argument in other propositions.

Most studies focus on the changes in propositional idea density of English texts. In the current study, we aimed to explore whether the findings from previous studies apply to Dutch written texts as well. The present study is exploratory in the sense that we try to gain insight into the propositional language behavior over time of healthy and Alzheimer's patients with only a limited amount of test material, i.e. the literary work of two renowned authors. A longitudinal approach was applied to study the changes in the propositional idea density in their work over time. For this literary focused research, novels from Claus and Elsschot were analyzed and their changes in propositional idea density scores over time were compared.

Measuring the propositional idea density of large texts and securing consistency throughout the propositional analysis of all texts requires a computerized approach. To the best of our knowledge, there is no manual on propositional analysis of Dutch texts, not to mention a computerized tool to execute the analysis. Consequently, we aimed to develop a computerized propositional idea density tool that is able to accurately measure the propositional idea density score of Dutch texts.

Our article is structured as follows. First, we introduce the concept of propositional idea density, discuss cognitive studies from previous research, and describe our efforts in creating a tool for its computational analysis in Section 2. Additional guidelines for the computation of propositional idea density in Dutch are presented in Appendix A. Section 3 holds the description of our experimental setup, including our method, materials, statistical analysis and results. A discussion of our results can be found in Section 4 after which we formulate a conclusion in Section 5.

2. Propositional Idea Density

Turner and Greene (1977) distinguish three types of propositional relations. For the predicate propositions, there is usually a verb that expresses an action or state, in relation to its argument(s). The second group includes negation words and quantifiers, which are examples of modifier propositions. Lastly, connective propositions serve as the connection between other propositions or facts expressed in the text. The number of propositional ideas within a text is defined by the sum of all occurences of these types of propositional relations within the text. Unlike simple linguistic measures such as word or sentence length, propositional idea density serves as a useful measure to analyze semantic changes in language, since it reflects the extent to which an individual connects different ideas instead of merely referring to different entities (Engelman et al. 2010).

A pioneering study within the subject of propositional idea density, that continues to inspire several other studies, is the Nun Study. The Nun Study was conducted by Snowdon et al. (1996) and investigated the effects of linguistic ability in early life on the cognitive function and development of AD in late life. The study is based on the assumption that developing AD or other dementias is associated with a lower cognitive reserve. The cognitive reserve hypothesis holds that individuals with a lower cognitive reserve are at greater risk to develop AD. Individuals with a higher cognitive reserve can develop AD as well, but the cognitive reserve compensates for the damage (i.e. they manifest symptoms of AD at a later stage in life and less severely) (Engelman et al. 2010). To measure the early-life linguistic ability, Snowdon et al. (1996) analyzed written autobiographies of 93 members of convents at the age of 22. For each text, they measured the propositional idea density. Following the assumption that a text is composed of one or more propositional units, the propositional idea density measure was defined as the average number of propositions per ten words.

As illustrated by (Snowdon et al. 2000), the sentence "*[[I was **born** [**in** Eau Claire, Wis.], [**on** May 24, 1913]] **and** [was **baptized** [**in** [**St. James** Church]]]].*" consists of 18 words and 7 ideas. The propositional units are divided by square brackets and the words in bold carry the propositional ideas. The number of ideas averaged per ten words results in a propositional idea density of 3.9. This measure is related to vocabulary, general knowledge and educational level, and reflects the cognitive reserve of an individual. Participants with a propositional idea density in the bottom third of the distribution were classified as low in propositional idea density (Snowdon et al. 1996). Next, Snowdon et al. examined the cognitive function of all participants. There are approximately 58 years between the age of their first writings and the cognitive tests. Additional neuropathological evaluation was carried out for the participants who were deceased. Snowdon et al. (1996) reported a strong connection between low propositional idea density in early life and low cognitive test scores in late life. Furthermore, 90 percent of all participants that developed AD had low propositional idea density scores in early life, whereas low propositional idea density scores were only present in 13 percent of the participants without AD. Since all sisters from the convents were exposed to the same lifestyle and environmental risk factors, they concluded that linguistic ability in early life is a good indicator for poorer cognitive function and AD in late life.

Snowdon et al. (2000) found supporting evidence in a subsequent study with 74 members of the convents, who were deceased. The women that met the neuropathological criteria for AD showed lower propositional idea density scores for their autobiographies in early life, as compared to the women who did not meet the criteria. In 2005, Riley et al. investigated the extent to which early life linguistic ability (i.e. propositional idea density) is related to the severity of cognitive decline or neuropathology in late life. They studied the handwritten autobiographies of 39 deceased members of the convents of whom 3 percent had intact cognitive functions, 25 percent had suffered from Mild Cognitive Impairment[1], 13 percent had suffered from Global Impairment[2], and 59 percent were diagnosed with dementia[3]. The results showed that low propositional idea density in early life was very strongly related to AD in late life, and significantly related to MCI in late life. They suggested that it is possible to identify individuals who are at risk of developing cognitive states with impaired memory functions (i.e. MCI and AD) in late life by measuring propositional idea density, as an indicator of linguistic ability, in early life.

Findings from the Nun Study gave rise to several related studies. Engelman et al. (2010) tested the same hypothesis, namely that early life low propositional idea density is a predictor of AD in late life, but on a different group of participants, from the Johns Hopkins Precursors Study. The participants from this Precursors Study differed from the Nun Study in gender, educational level and occupation. Engelman et al. (2010) measured the propositional idea density of written statements of 18 participants at the age of 22 who were later diagnosed with clinical AD, and 36 healthy controls, matched on age and gender. Different from the manual approach in the Nun Study, Engelman et al. used the Computerized Propositional Idea Density Rater (CPIDR 3) to measure the propositional idea density of all texts. Compared to the Nun Study, the differences in mean propositional idea density were considerably smaller, but still significantly lower for Alzheimer's patients than for the healthy controls. These results confirmed that individuals with a higher propositional idea density score in early life were less likely to develop AD in late life. However, it should be noted that the patients were diagnosed with AD following the ICD-9 guidelines, instead of a definitive post-mortem neuropathological diagnosis. In addition, Engelman et al. suggested that later developments of Alzheimer's disease for the healthy controls would restrict the propositional idea density associations to early-onset AD.

As an extension of previous studies, Farias et al. (2012) tested whether propositional idea density in late life continues to predict the trajectory of cognitive change, thus acting as a marker for cognitive

1. At least 1 specific area of impaired cognitive function, but intact global cognitive ability and daily living abilities.
2. Impaired global cognitive ability and/or daily living abilities, possible other impaired cognitive functions.
3. Impaired in memory and at least 1 other area of cognition, social and daily function impairment, decline in function from a previous level.

reserve. Their research relied on the study of (Kemper et al. 2001) that reported only a modest change in propositional idea density between young adulthood and old age. Farias et al. showed that higher propositional idea density scores predicted a slower rate of decline in global cognition, semantic memory, episodic memory and spatial abilities. Lower propositional idea density scores were associated with a steeper decline in cognitive function. The association between the rate of cognitive decline and propositional idea density was stronger for patients with MCI as compared to Alzheimer's patients, suggesting that this association is most apparent in the early stages of AD.

Spencer et al. (2012) carried out a longitudinal study to examine the stability of propositional idea density as an indicator and predictor of language decline with aging. They studied the variability of propositional idea density over a 16-year period for 127 middle-aged women who were part of the Australian Longitudinal Study of Women's Health (ALSWH). The age of the participants at the final survey ranged between 56 and 61, so they expected no change in propositional idea density due to cognitive decline. The researchers reported little change over time within the group. However, they observed much within-subject variation. This variation was attributed to the variability in text-length of the samples, rather than true propositional idea density changes for a subject. In other words, small sample sizes greatly reduce the precision of propositional idea density scores. In response to the problematic individual variation in propositional idea density, Ferguson et al. (2014) conducted a subsequent study that comprised 37,705 written samples from 19,512 healthy Australian women. In this second effort to gain insight into the range of normal or typical propositional idea density, they examined the change in language of women from three different age-groups over a period of 16 years. Ferguson et al. reported a relatively stable propositional idea density score across the young and mid adult lifespan, and a small but significant decline in propositional idea density for the oldest age group. However, the propositional idea density within each age group tended to increase over time. The problematic within-subject variability as reported in their previous study (Spencer et al. 2012) was reduced for samples that contain at least 100 words.

2.1 A Computerized Propositional Idea Density Tool for Dutch

Manual propositional annotation is very time consuming and offers no guarantee of consistency throughout the annotation task. For these reasons, we have developed an automatic propositional idea density tool for Dutch texts. This computerized tool enables us to measure the propositional idea density of large texts, such as those of the literary experiment described in this paper. The idea behind an automatic propositional analysis originates from a computer program called Computerized Propositional Idea Density Rater (CPIDR) (Brown et al. 2008). However, there is a lack of documentation regarding Dutch propositional analysis. As a result, in this paper, the propositional conditions that define the propositional idea density scores are based upon different articles and manuals concerning English propositional idea density (Brown et al. 2008, Chand et al. 2010, Chand et al. 2012), combined with insights in Dutch grammar (Vandeweghe 2013) to accurately cover most aspects of propositional units in Dutch.

Since the proposition count is based on the part-of-speech tags of each word, we first applied the part-of-speech tagger from the language tool Frog (Van den Bosch et al. 2007). Frog[4] is an integration of memory-based natural language processing modules for Dutch, developed by research groups in computational linguistics at Tilburg University and the University of Antwerp. For each word token[5], Frog returns the part-of-speech tag[6] and the lemmatized word. Following Brown et al. (2008), words with a particular part-of-speech tag were by default always counted as propositions (e.g. possessive pronouns, genitive articles, relative pronouns, etc.) or never counted as propositions (e.g. interjections, reflexive pronouns, interrogative pronouns, nominal adjectives, etc.). For words belonging to other word classes, more extended propositional conditions were applied to determine

4. Available on https://languagemachines.github.io/frog/.
5. The NLTK-RegexpTokenizer was used to tokenize all sentences.
6. Based on the CGN (Corpus Spoken Dutch) tagset.

	Elsschot	Claus	Total
Number of words	1,814	2,101	3,915
Number of propositions (Manual)	984	1,031	2,015
Number of propositions (Computerised)	978	1,020	1,998
False positives	28	32	60
Due to incorrect tag:	22	21	43
False negatives	34	43	77
Due to incorrect tag:	17	34	51

Table 1: Evaluation of the performance of the computerised propositional idea density rater for Dutch against manual annotation of literary works. A limited amount of sentences were selected from each novel.

whether we should count them as propositions or not. The detailed part-of-speech information (i.e. word class, type, inflection, genus, number, case, position, status, etc.) enabled us to create propositional conditions that are mainly restricted to the preceding and following tags in a sentence. A detailed description of the propositional conditions for Dutch are shown in Appendix A. Next, we divided the proposition count by the total number of words to measure the propositional idea density of each text studied.

It should be emphasized that the propositional idea density rater for Dutch is heavily based on rules that were designed for English texts. All basic concepts were translated to Dutch propositional rules and we have tried to implement the same way of thinking for all uncertain cases with no analogy in English. To evaluate whether the rule-based computerized tool was functioning properly as we intended, we compared the propositional selection against a sample of manual propositional annotation. It should be noted that the person who developed the tool is the same person that performed the manual annotation.

We manually annotated a limited number of sentences from all works of Elsschot and Claus. These sentences were randomly selected. The manual annotation chack resulted in a precision and recall for the system of 0.970 and 0.962, respectively. The error can be attributed to the fact that the sentences in these literary works tend to deviate from the standard sentence structures. Most errors occurred as a result of incorrect part-of-speech tags (See Table 1). Excluding these errors resulted in a precision score of 0.991 and recall score of 0.987. Overall, we can conclude that the computerized propositional idea density rater for Dutch performs well enough, according to our own guidelines, to use its results in propositional analysis.

In addition to the propositional idea density measure, we created a computerized baseline measure. The baseline score is defined by the total amount of verbs, adjectives, adverbs and prepositions, averaged over the total number of words. The idea behind the baseline score is that propositions mainly consist of verbs, adjectives, adverbs and prepositions. We aim to explore whether the baseline score performs well enough to function as a simplified propositional measure. In order to answer this question, we will examine the extent to which the baseline score correlates with the propositional idea density score of the texts.

3. Experiment

3.1 Method

We measured the changes in propositional idea density over a period of approximately 40 years. In order to conduct the propositional analysis, we collected the works of two Flemish writers. One of the authors, Hugo Claus, was diagnosed with AD. The second author, Willem Elsschot, has not been diagnosed with AD, nor has there been any reason to believe that he suffered from some type of

Title	Age (years)	Year	Number of words
De Metsiers	21	1950	31,644
De Hondsdagen	24	1953	43,215
De Koele Minnaar	27	1956	51,143
De Verwondering	33	1962	82,730
Omtrent Deedee	34	1963	32,683
Schaamte	43	1972	32,331
Het Jaar van de Kreeft	43	1972	53,454
Jessica!	48	1977	21,507
Het Verlangen	49	1978	54,350
Het Verdriet van België	54	1983	230,815
Een Zachte Vernieling	59	1988	39,592
De Zwaartvis	60	1989	21,882
Belladonna	65	1994	73,881
De Geruchten	67	1996	83,877
Een Slaapwandeling	71	2000	12,037
		Total:	865,141

Table 2: Novels written by Claus, with corresponding age, year of publication and the total number of words.

dementia. Based on previous studies (Engelman et al. 2010, Riley et al. 2005, Snowdon et al. 1996, Snowdon et al. 2000), we hypothesize that there will be differences in terms of propositional level in early life. We expect low propositional idea density scores in the early writings of Claus, as compared to Elsschot, since Claus performed poorly on cognitive tests in late life. Following Ferguson et al. (2014), chances are that the propositional idea density within the timespan of Elsschot's writings will increase over time. We do not expect to observe such an increase for the writings of Claus.

3.2 Materials

Hugo Claus is a Flemish author, born in 1929. He started writing novels when he was 22 years old and at the age of 75, he resigned from any form of writing, except for poetry. In 2006, Claus was officially diagnosed with AD. His diagnosis was communicated via email by his wife, Veerle (Ammerlaan 2008). Two years later, he submitted a request for euthanasia and died that same year. As reported by Van Dun (2013), Claus already suffered from disease-related complaints in the years before his diagnosis. In response to his word-finding and memory difficulties, he conducted a Mini Mental State Exam (MMSE). However, his MMSE score of 26 to 30 still fell into the 'normal' category. A CT scan showed cortical atrophy (i.e. dying of brain tissue), which was not disturbingly abnormal at the age of 74. In 2004, his cognitive state deteriorated and neuropsychological assessments revealed probable AD. In 2006, Claus' MMSE score had dropped to 19 and he had a score of 5.5 to 10 on the Hierarchic Dementia Scale. These scores, among others, resulted in the diagnosis of mild to moderate Alzheimer's disease (Van Dun 2013). To find out whether his linguistic ability changed throughout his life as a result of AD, we measured the propositional idea density score for all novels that Claus has written (see Table 2). His first novel was written at the age of 22 and he was 71 years old when his last novel was published. This last novel dated from six years before he was diagnosed with AD. The broad collection of works offers the possibility to investigate the language change up to a couple of years before the severe manifestation of his disease.

The second Flemish author, Willem Elsschot, was born in 1882. Elsschot represents the healthy control in our study, since there are no reasons to believe that he suffered from Alzheimer's disease-

Title	Age (years)	Year	Number of words
Villa des Roses	28	1910	43,298
Een Ontgoocheling	32	1914	16,293
De Verlossing	39	1921	33,241
Lijmen	41	1923	45,117
Kaas	51	1933	26,767
Tsjip	52	1934	27,939
Pensioen	55	1937	21,200
Het been	56	1938	19,720
De Leeuwentemmer	58	1940	22,101
Het Tankschip	59	1941	13,395
Het Dwaallicht	65	1947	13,778
		Total:	282,849

Table 3: Novels written by Elsschot, with corresponding age, year when finished writing and the total number of words.

related complaints. The novels for the propositional analysis were obtained from his collected work (Elsschot 1980). An overview of Elsschot's writings is shown in Table 3. For each novel, we used the year that was mentioned at the bottom of the text, which does not always match its year of publication. Elsschot wrote his first novel at the age of 28 and his last novel dated from 1947, when he was 64 years old. Although this time span is slightly shorter than Claus', it still offers a representative collection of works throughout Elsschot's writing career.

To avoid any influence of differences in length, we limited our propositional analysis of each novel until the sentence that contained the 12.030th word, which is the same as the number of words of the shortest novel (i.e. Claus' *Een slaapwandeling*).

3.3 Statistical Analysis

For the statistical analysis of the experiment, we employed a linear regression model. Propositional idea density was used as a continuous dependent variable, while the independent variables consisted of author (categorical, with two levels) and authors' age (continuous, in years). Thereafter, we repeated the linear regression approach but replaced the propositional idea density score by the simplified baseline score. The performance of the baseline measure as compared to the elaborate propositional idea density measure was established by a Spearman correlation analysis.

3.4 Results

A linear regression model accounted for 74.5 percent of the variation in propositional idea density. For the writings of Elsschot, we observed a significant increase in propositional idea density (β = 0.0004, t(25) = 2.68, $p < 0.05$). As a result, the propositional idea density at the end was significantly higher compared to the estimated propositional idea density at starting point (t(25) = 61.76, $p <$ 0.001). On the other hand, Claus' decrease in propositional idea density was significantly different from Elsschot's increase in propositional idea density (β = - 0.0006, t(25) = -3.74, p = 0.001). Figure 1 illustrates this steep increase in propositional idea density for the writings of Elsschot as compared to the slight decrease in Claus' propositional idea density over time. There was no significant difference in propositional idea density between the writings of Elsschot and Claus at the age of 28 (t(25) = -0.53, p = 0.599).

Next, we examined the density ellipse of the bivariate normal distribution of the baseline score and the propositional idea density score. As determined by a significant Pearson correlation coefficient

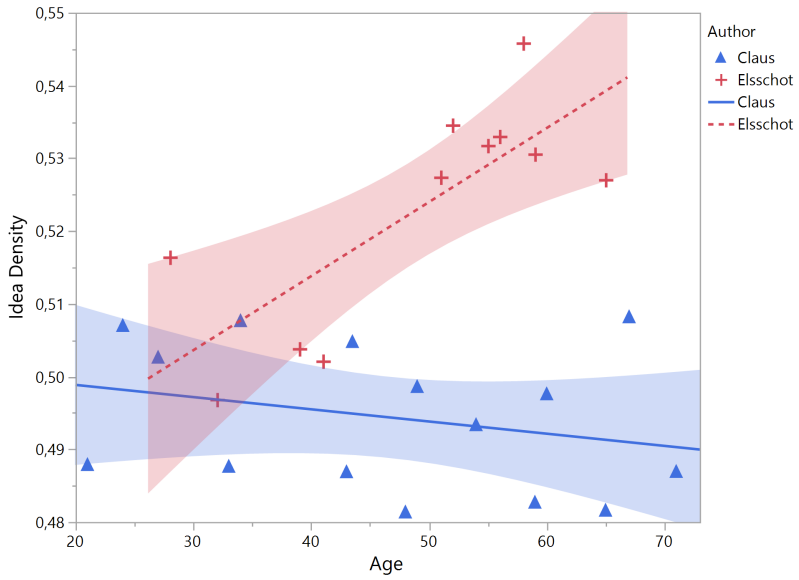

Figure 1: Linear regression lines of mean propositional idea density scores of Claus and Elsschot over time. 95% confidence bands are added around the regression lines.

of 0.889 ($p < 0.001$), it appears that there is a strong correlation between both scores. The non-parametric Spearman's rank correlation coefficient of 0.859 ($p < 0.001$) confirmed the previous statement that there is a strong correlation between the propositional idea density score and the simplified baseline score of large texts. For this reason, we compared previous results with regard to propositional idea density to the linear regression model of the baseline score. In line with propositional idea density, the baseline score of Elsschot's writings as estimated in late life was significantly higher as compared to the estimated baseline measure at starting point ($t(25) = 57.09$, $p < 0.001$). Claus' estimated baseline score at the age of 28 was not significantly lower than Elsschot's baseline score ($t(25) = -0.71$, $p = 0.488$). The change in baseline score over time for the writings of Claus differed significantly from the change in Elsschot's writings ($\beta = -0.0005$, $t(25) = -2.88$, $p < 0.05$). In contrast with changes in propositional idea density, we observed no significant increase in baseline score for the writing of Elsschot over time ($\beta = 0.00019$, $t(25) = 1.18$, $p = 0.250$).

4. Discussion

The findings from this experiment provide support for the hypothesis that the propositional idea density score of healthy individuals increases over time. Furthermore, Claus' change in propositional idea density over time differed significantly from the change observed for the writings of Elsschot. In contrast to the steep increase in propositional idea density of Elsschot's writings, we found a slight decrease in propositional idea density for the writings of Claus. These findings are consistent with the findings from Ferguson et al. (2014).

Our hypothesis prior to analysing the data stated that an observable difference in propositional idea density in early life between Claus and Elsschot was expected. Compared to Elsschot, no significant difference in propositional idea density score of Claus' writing estimated at the age of 28 was seen after examining the linear regression analysis. A significant difference in propositional idea density between the authors is more pronounced in late life, as the propositional idea density of Claus slightly decreases and the propositional idea density of Elsschot greatly increases over time.

Overall, our findings provide no support for the assumption that early life linguistic ability (i.e. propositional idea density) is related to cognitive (dys)function in late life (Snowdon et al. 1996, Snowdon et al. 2000, Riley et al. 2005, Engelman et al. 2010). However, given that we carried out a case study comparing only one author that eventually developed AD to one author that did not, the question regarding the representativeness of both authors as reflecting the propositional behavior of healthy individuals and Alzheimer's patients remains to be answered.

The results of the manual propositional annotation answered the question of whether it is possible to create a computerized tool to measure the propositional idea density of Dutch texts. In view of the high precision and recall scores, we were able to properly measure the propositional idea density of large texts by using the computerized tool.

Next, we aimed to explore whether a simple baseline measure could function as an appropriate simplified propositional idea density measure for large texts. For the novels used in the experiment, correlation results indicated a strong relationship between propositional idea density and the baseline score. These findings suggest that a simplified propositional measure (i.e. baseline score) results in a similar pattern as the elaborate propositional idea density score. Nevertheless, a subsequent linear regression analysis revealed no significant increase in baseline score for the writings of Elsschot. Therefore, we conclude that propositional idea density may be a more precise measure to establish changes in linguistic ability over time. Further studies should elaborate further on propositional idea density as a more precise measure by studying the differences in types of propositions between and within the authors' writings.

4.1 Limitations

Our case study consisted of only one author representing the population of healthy controls and one author representing the population of Alzheimer's patients. Whether these findings account for the propositional behavior of a larger population cannot be determined on the basis of the present data. This study would also benefit from more balanced data on age and year of publication, since novels from both authors were not published on an equal and annual basis. Furthermore, the last work of Elsschot was published around the time of Claus' first novel. As language changes, propositional idea density might change as well. The effect of language change due to time of writing on propositional idea density has not been accounted for so far and remains unclear. Differences in propositional idea density between both authors might also be caused by differences in genre. There is a plausible effect of the novels' genre on propositional idea density that has, however, not been focused on in the present study. Another limitation to the data of this experiment is the lack of authors with MCI. Including the changes in propositional idea density of authors with MCI might be very useful to determine the position of MCI as compared to healthy elderly and Alzheimer's patients.

A linear regression model was used as a default method to analyse the results. This does not mean that a linear regression model is necessarily the best fit for this dataset. Further studies should compare the use of different models on a similar dataset.

Whereas previous studies (Snowdon et al. 1996, Kemper et al. 2001, Engelman et al. 2010, Farias et al. 2012, Chand et al. 2012, Bryant et al. 2013, Ferguson et al. 2014) investigated the effect of educational level on propositional idea density, no such effect was tested in the present study. Due to incomplete data, we decided to leave the possible effect of educational level out of the experiment. For the experiment, there were only two participants, which makes it impossible to accurately measure the effect of educational level on propositional idea density. Furthermore, given the difference in time, we cannot simply compare the levels of education. Thus, regardless of the possible effect on propositional idea density as reported in previous studies, no such differences are covered in the present study.

Overall, we managed to develop a computerized propositional count for Dutch texts. However, there are still many shortcomings. First, the data that was used during the (handcrafting) development of the computerized tool is the same as the data used for the experiment. Consequently, the

computerized tool might perform worse when applied to different texts. Furthermore, the evaluation of the computerized tool was performed by the same person who developed the tool. Such evaluation on performance offers less certainty that the propositional conditions are reliable. The computerized propositional idea density rater for Dutch texts would greatly benefit from tests by multiple human raters. However, the inter- and intra-rater reliability would still suffer from the lack of documentation regarding propositions in Dutch. Second, the performance of the computerized tool strongly depends on the accuracy of the part-of-speech tagger. Most of the miscounted propositions were a result of falsely tagged words. To avoid such miscounts, the performance of the part-of-speech tagger needs to be improved. Third, there are several limitations with regard to the complexity of language. So far, the computerized tool has not been able to detect sentence structures that present idioms or metaphors. While humans have little trouble recognizing idioms and metaphors, they pose more problems for the computerized tool, which treats such structures as regular clauses. Other instances of sentences that are difficult to manage are very long sentences or sentences that strongly deviate from standard sentence structures.

Multiple limitations of the computerized tool are due to the complexity of Dutch language in particular. For instance, the adjustment rules that comprise the computerized tool do not include all possible occurrences of replacement auxiliary verbs. Such replacement verbs can function as lexical verbs as well as auxiliary verbs, but the recognition of the appropriate function in the sentence requires human intuition. For example, the verb *loopt* in "Alles loopt mis" (*Everything goes wrong.*) functions as a replacement copula verb, followed by an adjective. Theoretically, copula verbs that are accompanied by an adjectival or adverbial phrase should not be counted as a separate proposition. However, this restriction would mistakenly refrain from counting the lexical verb *loopt* in "Zij loopt snel" (*She runs fast.*) as a proposition. Another limitation is related to the Dutch typological system. In addition to the SVO (i.e. Subject-Verb-Object) word order as used in English, Dutch sentences can follow a SOV-order (i.e. Subject-Object-Verb) as well. As a result, many words can be placed between the subject and main verb, or between the auxiliary verb and lexical verb. The more words that are placed between two verbs, the more difficult it becomes to properly identify verbs as being auxiliary verbs or lexical verbs.

To conclude, further studies involving more age-matched participants from each cognitive group are needed to confirm our findings and to allow more definite conclusions. Future studies should also examine the effect of educational level and text genre on propositional idea density. Finally, It will be important to create more clear guidelines concerning propositions in Dutch in order to adjust the computerized tool and to evaluate its performance.

5. Conclusion

In this study, we aimed to explore whether there are differences in propositional idea density between healthy individuals and patients with Alzheimer's disease. We hypothesized that the propositional idea density in early life would be significantly lower for Alzheimer's patients as compared to healthy elderly. We expected the propositional idea density of healthy individuals to increase over time. No such increase was expected for Alzheimer's patients. With regard to the propositional idea density in late life, we hypothesized that this measure would be higher for healthy elderly as compared to patients with AD.

Prior to the experiment, we developed a computerized tool that measures the propositional idea density of Dutch texts and secures a consistent and time-saving propositional analysis. Following the existing guidelines and reports concerning propositions in English texts, we created a part-of-speech-based propositional idea density measure for Dutch that performs with high precision and recall. After looking at the errors of the evaluation against a human annotator, we concluded that the computerized tool performed well enough to be used for the propositional analysis of large texts.

To establish the difference in propositional idea density change between healthy elderly and Alzheimer's patients, a longitudinal experiment comparing the changes in propositional idea density

between two renowned authors was carried out. On the one hand, the results did not support the hypothesis that the propositional idea density of the author with AD (i.e. Claus) in early life (at the age of 28) would be significantly lower than the propositional idea density of the healthy author (i.e. Elsschot). On the other hand, our findings support the assumption that propositional idea density of healthy elderly significantly increases over time. This increase in propositional idea density significantly differed from the decrease in propositional idea density of the author with AD.

We checked whether similar changes and differences were observed with a simplified propositional measure, referred to as the baseline score, as compared to propositional idea density. A similar pattern between both measures would indicate that the differences in propositional idea density between healthy elderly and Alzheimer's patients can be established by applying a more accessible measurement. Regardless of the correlation between both measures in the experiment, Elsschot's changes in baseline measure deviated from the changes in propositional idea density. As a result, the more precise propositional idea density measure was preferred over the simplified baseline score.

We concluded this study by emphasizing several limitations to the study and by suggesting a number of recommendations for future studies. Overall, the small sample size presents the biggest limitation of the present study. With this in mind, we suggested to take on further studies to obtain a better understanding of differences in propositional idea density between healthy individuals and Alzheimer's patients. Despite the large amount of studies that examine propositional idea density as a possible predictor of AD, current results should be considered preliminary with regard to propositional idea density of Dutch texts. This study calls for further research regarding the understanding of propositions in Dutch. The guidelines, provided in Appendix A, might open the way to further concretize the concept of propositions in Dutch.

Overall, findings from this study do not reject the idea that propositional idea density functions as an indicator of Alzheimer's-related cognitive decline in late life. To that end, a computerized propositional idea density rater greatly facilitates the measurement of propositional idea density.

References

Ammerlaan, Robbert (2008), *De laatste van mijn demonen*, Uitgeverij De Bezige Bij b.v.

Brown, Cati, Tony Snodgrass, Susan J Kemper, Ruth Herman, and Michael A Covington (2008), Automatic measurement of propositional idea density from part-of-speech tagging, *Behavior research methods* **40** (2), pp. 540–545, Springer.

Bryant, Lucy, Elizabeth Spencer, Alison Ferguson, Hugh Craig, Kim Colyvas, and Linda Worrall (2013), Propositional idea density in aphasic discourse, *Aphasiology* **27** (8), pp. 992–1009.

Chand, Vineeta, Kathleen Baynes, Lisa Bonnici, and Sarah Tomaszewski Farias (2010), Analysis of idea density (aid): A manual, *University of California at Davis*.

Chand, Vineeta, Kathleen Baynes, Lisa Bonnici, and Sarah Tomaszewski Farias (2012), A rubric for extracting idea density from oral language samples, *Current protocols in neuroscience* pp. 10–5, Wiley Online Library.

Croisile, Bernard, Bernadette Ska, Marie-Josee Brabant, Annick Duchene, Yves Lepage, Gilbert Aimard, and Marc Trillet (1996), Comparative study of oral and written picture description in patients with alzheimer's disease, *Brain and language* **53** (1), pp. 1–19, Elsevier.

Elias, Merrill F, Alexa Beiser, Philip A Wolf, Rhoda Au, Roberta F White, and Ralph B D'agostino (2000), The preclinical phase of alzheimer disease: a 22-year prospective study of the framingham cohort, *Archives of neurology* **57** (6), pp. 808–813, American Medical Association.

Elsschot, Willem (1980), *Verzameld Werk*, Em. Querido's Uitgeverij B.v.

Engelman, Michal, Emily M Agree, Lucy A Meoni, and Michael J Klag (2010), Propositional density and cognitive function in later life: Findings from the precursors study, *Journals of Gerontology Series B: Psychological Sciences and Social Sciences* **65** (6), pp. 706–711, Oxford University Press.

Farias, Sarah Tomaszewski, Vineeta Chand, Lisa Bonnici, Kathleen Baynes, Danielle Harvey, Dan Mungas, Christa Simon, and Bruce Reed (2012), Idea density measured in late life predicts subsequent cognitive trajectories: Implications for the measurement of cognitive reserve, *Journals of Gerontology Series B: Psychological Sciences and Social Sciences* **67** (6), pp. 677–686, Oxford University Press.

Ferguson, Alison, Elizabeth Spencer, Hugh Craig, and Kim Colyvas (2014), Propositional idea density in women's written language over the lifespan: Computerized analysis, *Cortex* **55**, pp. 107–121, Elsevier.

Hamilton, Heidi Ehernberger (1989), *Conversations with an Alzheimer's patient: an interactional examination of questions and responses*, Georgetown University.

Kemper, Susan, Lydia H Greiner, Janet G Marquis, Katherine Prenovost, and Tracy L Mitzner (2001), Language decline across the life span: findings from the nun study., *Psychology and aging* **16** (2), pp. 227, American Psychological Association.

Kintsch, Walter and Janice Keenan (1973), Reading rate and retention as a function of the number of propositions in the base structure of sentences, *Cognitive psychology* **5** (3), pp. 257–274, Elsevier.

Kintsch, Walter and Teun A Van Dijk (1978), Toward a model of text comprehension and production., *Psychological review* **85** (5), pp. 363, American Psychological Association.

Kluger, Alan, Steven H Ferris, James Golomb, Mary S Mittelman, and Barry Reisberg (1999), Neuropsychological prediction of decline to dementia in nondemented elderly, *Journal of geriatric psychiatry and neurology* **12** (4), pp. 168–179, Sage Publications Sage CA: Thousand Oaks, CA.

Le, Xuan, Ian Lancashire, Graeme Hirst, and Regina Jokel (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), pp. 435–461, Oxford University Press.

Petersen, Ronald C, James C Stevens, Mary Ganguli, Eric George Tangalos, J L Cummings, and S T DeKosky (2001), Practice parameter: Early detection of dementia: Mild cognitive impairment (an evidence-based review) report of the quality standards subcommittee of the american academy of neurology, *Neurology* **56** (9), pp. 1133–1142, AAN Enterprises.

Riley, Kathryn P, David A Snowdon, Mark F Desrosiers, and William R Markesbery (2005), Early life linguistic ability, late life cognitive function, and neuropathology: findings from the nun study, *Neurobiology of aging* **26** (3), pp. 341–347, Elsevier.

Snowdon, David A, Lydia H Greiner, and William R Markesbery (2000), Linguistic ability in early life and the neuropathology of alzheimer's disease and cerebrovascular disease: Findings from the nun study, *Annals of the New York Academy of Sciences* **903** (1), pp. 34–38, Wiley Online Library.

Snowdon, David A, Susan J Kemper, James A Mortimer, Lydia H Greiner, David R Wekstein, and William R Markesbery (1996), Linguistic ability in early life and cognitive function and alzheimer's disease in late life: Findings from the nun study, *Jama* **275** (7), pp. 528–532, American Medical Association.

Spencer, Elizabeth, Hugh Craig, Alison Ferguson, and Kim Colyvas (2012), Language and ageing–exploring propositional density in written language–stability over time, *Clinical linguistics & phonetics* **26** (9), pp. 743–754, Taylor & Francis.

Tarawneh, Rawan and David M Holtzman (2012), The clinical problem of symptomatic alzheimer disease and mild cognitive impairment, *Cold Spring Harbor perspectives in medicine* **2** (5), pp. a006148, Cold Spring Harbor Laboratory Press.

Turner, Althea and Edith Greene (1977), *The construction and use of a propositional text base*, Institute for the Study of Intellectual Behavior, University of Colorado Boulder.

Van den Bosch, Antal, Bertjan Busser, Sander Canisius, and Walter Daelemans (2007), An efficient memory-based morphosyntactic tagger and parser for dutch, *LOT Occasional Series* **7**, pp. 191–206, LOT, Netherlands Graduate School of Linguistics.

Van Dun, Kim (2013), *A neurolinguistic analysis of the evolution of written and oral language of an author with dementia of the alzheimer's type*, Master's thesis, Vrije Universiteit Brussel.

Vandeweghe, Willy (2013), *Grammatica van de Nederlandse zin*, Garant.

Appendix A. Propositional Conditions

These examples are based on the Corpus Gesproken Nederlands (CGN)-tagset (Van Eynde 2004).

Always count as proposition:

- Genitive articles
  E.g. *des* duivels, *der* Nederlandse taal, de kracht *ener* vrouw
- Melted prepositions
  E.g. *ten* strijde, *ten* hoogste, *ter* plaatse
- Adjectives (except nominal adjectives)
  E.g. een *mooi* huis, het *grootste* paard, die stok is *lang*
- Numerals (except nominal numerals)
  E.g. *2, drie*, de *vierde* man
- Possessive pronouns
  E.g. *jouw* vriend, *d'r* paard, *zijne* excellentie
- Personal pronouns: Genitive
  E.g. *mijns* gelijke, velen *onzer*
- Reciprocal pronouns: Genitive
  E.g. *elkaars, mekaars, elkanders*
- Relative pronouns
  E.g. de man *die* daar staat, het kind *dat* je ziet, *hetgeen* je daar ziet

Never count as proposition:

- Punctuation
  E.g. *, . ; ! ? ( )][-*
- Special characters (Names and unknown)
  E.g. comments, foreign words, meta data, incomprehensible words
- Interjections
  E.g. *uh, oei, amai*
- Final prepositions
  E.g. liep de trap *af*, dit komt wel eens *voor*
- All articles (except genitive articles)
  E.g. *het* kind, *de* bloemen, *een* man, op *den* duur
- Proper nouns & Type names (except genitive forms)
  E.g. die *stoel*, de *Middeleeuwen, Linux, huisjes*
- Nominal adjectives
  E.g. de *rijken*, ik sta in het *nat*, een *grote* met tartaar
- Nominal numerals
  E.g. iets aan *derden* verkopen, met z'n *vieren*, de *eersten*
- Reflexive pronouns
  E.g. *zich, zichzelf*

- Personal pronouns (Except genitives)
  E.g. *Ik, wijzelf, jou, me*

- Reflexive x Personal pronouns (they receive their own tag)
  E.g. *mij, onszelf, mezelf*

- Reciprocal pronouns (Except genitives)
  E.g. *elkaar, mekaar, elkander*

- Interrogative pronouns
  Only example: *Watte*?

- Exclamatory pronouns
  E.g. *Wat/Welk* een dwaasheid, *wat* kan jij liegen zeg

Word-specific conditions:

- Conjunctions:
  - '*dan*' preceded by '*als*' = 1 proposition, '*als*' is counted
  - '*dan*' preceded by '*meer*' = 1 proposition, '*meer*' is counted
  - Always count remaining conjunctions
    E.g. Jan *en* Peter, *omdat* ze naar huis ging

- Relative x interrogative pronouns:
  - Never count standard pronoun: (subject of sentence)
    E.g. *wie* gaat er mee?
  - Always count all remaining rel. int. pronouns
    E.g. *Wiens* hoed is dit, *waar* ga je naartoe, *welke* kinderen

- Demonstrative pronouns:
  - Never count: demonstrative standard pronouns
    E.g. *zulks, datte, die*
  - Never count: standard and dative nominal determiner
    E.g. *degene, genen*, dat is dan bij *dezen* beslist
  - Adverbial pronouns: always count, except '*er*'
    E.g. *hier, daar, d'r*
  - Always count: all remaining demonstrative pronouns
    E.g. *diens* voorkeur, *dat* boek, *deze* man, een *dezer* dagen

- Indefinite pronouns:
  - Indefinite pronominal standard pronoun: only count '*wat*'
  - Indefinite pronoun '*Geen*' followed by '*tenzij*' = 1 proposition, only '*tenzij*' is counted
  - Indefinite pronoun '*enkele*' preceded by '*geen*' = 1 proposition, only '*geen*' is counted
  - Indefinite pronoun '*meer*' followed by '*geen*' = 1 proposition, only '*geen*' is counted
  - Never count: genitive adverbial pronouns
    E.g. quantitative '*er*'
  - Always count: all remaining indefinite pronouns
    E.g. *andermans, (n)ergens, elk* huis, *geen* kind, in *mindere* mate, de *meest* gezochte

- Initial prepositions:
  - '*te*' followed by a verb = 1 proposition, verb is counted
  - '*om*' followed by '*te*' and a verb = 1 proposition, verb is counted
  - '*aan*' followed by '*het*' and a verb = 1 proposition, verb is counted
  - Always count: all remaining initial propositions
    E.g. *met* een lepeltje

- Adverbs:
  - '*zowel*' followed by '*als*' = 1 proposition, '*als*' is counted
  - '*nog*' followed by '*nooit*' = 1 proposition, '*nooit*' is counted
  - '*niet*' followed by '*meer*' = 1 proposition, '*meer*' is counted
  - '*niet*' followed by '*tenzij*' = 1 proposition, '*tenzij*' is counted
  - '*ook*' followed by '*al*' = 1 proposition, '*al*' is counted

- Verbs:
  For all verbs that are no auxiliary verbs:
  - Never count nominal past/present particles
    E.g. een *gekwetste*, het *resterende*, de *wachtenden*
  - Never count nominal infinitives, except when preceded by '*aan het*'
    E.g. (het) *schaatsen*, (het) *spelen*
  - Always count all remaining verbs

  For all forms of '*hebben*':
  - Always count lexical verb '*hebben*'
    E.g. ik *heb* één zus en twee broers, ik *heb* een jurk aan
  - Never count auxiliary verb '*hebben*'
    E.g. ik *heb* de was gedaan

  For all forms of '*zijn*' & '*worden*':
  - Copula verb followed by an adjectival phrase = 1 proposition, adjective is counted
    E.g. ik *ben* boos, mijn opa *wordt* oud
  - Copula verb followed by a noun phrase = 1 proposition, '*zijn*'/'*worden*' is counted
    E.g. ik *ben* een student, mijn zus *wordt* moeder
  - Never count auxiliary/passive verbs '*zijn*'/'*worden*'
    E.g. wij *zijn* getrouwd, zij *werden* gepest
  - Lexical verb '*zijn*' (zich bevinden) followed by prepositional/adjectival phrase = 1 proposition, preposition or adjective is counted
    E.g. moeder *is* in de keuken, zij *zijn* hier

  For all secondary copula verbs (*blijven, blijken, lijken, schijnen, dunken, heten, voorkomen*):
  - Always count lexical verbs:
    E.g. ik *blijf* hier, zij *lijkt* op haar moeder, morgen moet haar zaak *voorkomen*, mijn mama *heet* Els, de maan *schijnt* fel
  - Sec. copula verb followed by a prepositional/adjectival/verb phrase = 1 proposition, adjective/preposition/verb is counted
    E.g. zij *lijkt* zeer tevreden, zij *scheen* onder de indruk, hij *bleek* gebeten door een hond
  - Auxiliary verb '*blijf*' followed by a verb = 1 proposition, verb is counted
    E.g. ik *blijf* schrobben

- Sec. copula verb followed by a noun phrase = 1 proposition, sec. copula is counted
  E.g. Hij *blijkt* professor aan de hogeschool, haar lippen *leken* wel rozen

For all modal verbs (*zullen, moeten, willen, mogen, durven, hoeven*)
- Always count lexical verbs:
  E.g. *mag* ik mee?, ik *wil* een groot feest, ik *hoef* geen geld
- Never count modal verbs:
  E.g. Zij *moet* gaan slapen, ik *durf* niet aan te bellen

For replacement auxiliary verb '*gaan*':
- '*gaan*' followed by an infinitive verb = 1 proposition, inf. verb is counted
  E.g. Wij *gaan* winkelen
- Always count lexical verb '*gaan*'
  E.g. Zij *gaan* naar hun oma, ik *ga* op vakantie

For replacement auxiliary verbs *'zitten', 'liggen', 'staan', 'beginnen'*:
- Verbs followed by '*te*' and an infinitive verb = 1 proposition, inf. verb is counted
  E.g. hij *zit* heel de dag te zagen
- Always count lexical verbs
  E.g. moeder *ligt* in bed

Example sentences propositional analysis:
Each number represents a propositional unit. The word in bold is the element of the propositional unit that is counted.

Boeken vallen uit de kast.
(1) **Vallen**, boeken
(2) **Uit** de kast, vallen

De toestand is ernstig.
(1) Is **ernstig**, de toestand

Dit loopt verkeerd af.
(1) **Loopt** af, dit
(2) **Verkeerd** aflopen

We zien dat de papa aan het dutten is in de zetel.
(1) **Zien**, we
(2) **dat**, 1 3
(3) is aan het **dutten**, papa
(4) **in** de zetel