UNIVERSITY OF TORONTO

DEPARTMENT OF COMPUTER SCIENCE

M.SC. PAPER

Longitudinal Detection of Dementia Through Lexical and Syntactic Changes in Writing

Xuan LE

Supervisor: Graeme HIRST

22 January 2010

Longitudinal Detection of Dementia Through Lexical and Syntactic Changes in Writing

:: Xuan Le :: lexuan@cs.toronto.edu ::

Department of Computer Science
University of Toronto
Toronto, Ontario, M5S3G4

Abstract

Studies of language in dementia have concluded that, along with a general cognitive decline, linguistic features are also negatively affected. Studies of the language of healthy elders also observe a linguistic decline, but one which, in contrast, is markedly less severe than that induced by dementia. In this paper, we examine whether the disease can be detected from the diachronic changes in written texts and, more importantly, whether it can be clearly distinguished from normal aging. Lexical and syntactic analyses were conducted on fifty-one novels by three prolific literary authors: Iris Murdoch, P. D. James, and Agatha Christie. Murdoch was diagnosed with Alzheimer's disease shortly after finishing her last novel; James, at 89 years of age, continues to publish critically-acclaimed works; Christie, whose last few novels are deemed strikingly subpar compared to her previous works, presents an interesting case study of possible dementia. The lexical analysis reveals significant patterns of decline in Murdoch's and Christie's later novels, while James's rates remain relatively consistent throughout her career. The syntactic measures, though unveiling fewer significant linear trends, discover a cubic model of change in Murdoch's novels, with a deep decline around her 50s. Our findings provide further support for the hypothesis that dementia, which manifests clearly in lexical features, can be detected in writing.

Acknowledgements

I am grateful to my parents, Lê Đình Nho and Nguyễn Thị Hiền, who have always supported me and ensured that I have every opportunity to succeed, in school and in life; to my aunt Rachel and her husband Jeremy Harvey for their warmth and hospitality, despite the freezing cold and the occasional snow storms, during my stay in Canada; and to the rest of my family for sending their love and encouragement from almost every corner of the world.

I am indebted to my undergraduate instructors at the University of Toronto at Mississauga, especially Ms. Michelle Craig and Mr. Andrew Petersen of the Computer Science Department, for their dedication to teaching and personal attention to help students succeed. I thank Dr. Chris Koenig-Woodyard and Dr. Mark Crimmins of the English Department, and my high school instructor, Mr. Hoàng Ngọc Hùng, for instilling in me, a non-native speaker of English, a love and an appreciation for the language and its literature. It was the instructors' encouraging feedback that gave me the confidence to pursue graduate studies and research in Computational Linguistics at the University of Toronto.

I would like to express my gratitude to my graduate supervisor, Dr. Graeme Hirst, who entrusted me with this project, for his invaluable advice throughout my research career, and especially for taking the time from his New Year break to proofread multiple revisions of this paper. The end product would have been of a lesser quality, had it not been for the constructive feedback from Dr. Hirst, my second reader Dr. Suzanne Stevenson of the Computer Science Department, and Dr. Ian Lancashire of the English Department. Dr. Lancashire, who initiated the analysis of Agatha Christie's writings, has shared his vast knowledge of literature in support of the current study. This study has also benefited greatly from the clinical expertise and insights of Dr. Regina Jokel, of Baycrest's Research Centre for Aging and the Brain. In addition, I must acknowledge the generous financial support of Google through their Google Research Awards program, the Natural Sciences and Engineering Research Council of Canada, and the University of Toronto.

Finally, my sincere thanks to my high school, undergraduate and graduate friends who, virtually and in person, have never failed to remind me that there is more to life than exams and research!

Contents

List of Figures

List of Tables

1 Introduction

Alzheimer's disease, along with other types of dementia, is among the most prevalent geriatric conditions affecting a large proportion of the aging population. Clinical assessment of dementia involves several diagnostic procedures, which may be uncomfortable, tiresome, or even stressful for the individuals undergoing diagnosis. Recent research into dementia has demonstrated that the disease negatively affects the linguistic abilities of patients in both speech and writing. This fact presents the possibility of developing nonintrusive evaluation techniques that require minimal involvement from the patients, which may be used in conjunction with clinical assessments or on their own as an early detection tool.

However, that signs of decline exist is insufficient in itself to conclude a diagnosis, since an individual's language behaviour may change with advancing age regardless of his/her cognitive health. In addition, person-to-person variations in initial linguistic abilities require that the decline be examined in a longitudinal context on an individual basis. Several research groups in psycholinguistics have investigated the differences in linguistic change between diagnosed dementia patients and healthy elders, though mostly in a group study setting. Maxim and Bryan (1994) point out the drawbacks of such an approach:

> There is now a consensus that the language symptomatology in [dementia of the Alzheimer type] is heterogeneous [. . .]. One of the main problems in group studies that use a profile approach is that very little can be said about what the individual patient can and cannot do. Single case studies [. . .] have, to some extent, helped, particularly because they often point to specific deficits and dissociations between deficits that are unlikely to be found in group studies (p. 176).

Among the relatively few studies that pursued the case-study approach, Garrard et al. (2005) examined text samples drawn at random from three novels written by Iris Murdoch, an English author who was diagnosed with Alzheimer's disease shortly after the publication of her last novel in 1995. The research group pioneered in applying computational methods to literary works; most of the analysis, however, was performed on a sparse, minimal dataset that was insufficient to correctly represent Murdoch's linguistic levels. Conducting their analysis on a larger dataset, Lancashire and Hirst (2009) studied the lexical characteristics of sixteen novels by Agatha Christie, an author whose last few novels received largely negative reactions from critics—an unusual departure from her previous works—and whose biographies reveal a period of writing difficulties that she encountered while working on these novels. Lancashire and Hirst (2009) discovered a gradual decrease in vocabulary size over time, as well as an increase in repeated phrases and indefinite nouns (*thing*, *something*, and *anything*), which were particularly evident in Christie's later novels.

Proceeding further in this direction, in our research we sought writing samples from healthy elderly adults and diagnosed dementia patients to test the hypothesis that the disease manifests itself in language production, as well as writing samples from suspected, undiagnosed cases of dementia, to see whether linguistic indicators can serve as a diagnostic tool. Each set of texts must be written by the same individual, be of substantial length, and span several decades, from the writer's youth into his/her late-70s or 80s. While it is difficult to recruit participants in the general population who have enough writing samples preserved from before the digital age to meet our criteria, prolific literary authors provide us with a wealth of data for textual analysis. Following Garrard et al. (2005), we chose the writings of Iris Murdoch as the linguistic model of dementia patients and, in addition, contrasted it against the linguistic model of healthily aging adults represented by the writings of crime fiction author P. D. James. Expanding on the work by Lancashire and Hirst (2009), we also analyzed the novels of Agatha Christie, which present an interesting case study of possible undiagnosed dementia.

In this section, we provide an overview of the clinical background on dementia, as well as findings by other studies into language of the elderly, before formalizing our hypothesis.

1.1 Clinical Background on Dementia

Dementia is a clinical syndrome that can be caused by a range of diseases or injuries to the brain, pervasively found in the aging population.[1] Given the recent increase in life expectancy, dementia

[1] Clinical facts from this section, unless indicated otherwise, are drawn from Blazer and Steffens (2009).

has become “epidemic and one of the top 10 causes of disabilities in developed countries” (Blazer and Steffens, 2009, p. 243). Research studying the prevalence of dementia, conducted with sample populations from different communities, has suggested an estimate of up to 23% of the older population to be affected by the syndrome. A recent report predicts that, by 2038, the number of Canadians living with dementia will escalate from 500,000 in 2008 to 1.1 million, costing the public health-care system an estimated $153-billion, compared to the current $15-billion, per year.[2]

Dementia patients suffer from a marked decline in several cognitive abilities, which may include memory, orientation, and language comprehension and production, among other areas, causing profound impact on the patients’ day-to-day functioning. In contrast, although healthily aging adults may also experience a decline in their cognitive abilities, this age-related decline is significantly less severe. Alzheimer’s dementia is the most common form of dementia, accounting for 50–75% of all diagnosed cases. It is caused by Alzheimer’s disease (AD), a cognitive disorder estimated to affect nearly 10% of the population over age 65, and 25–40% of those over age 85. AD is often characterized by an insidious onset and a progressive impairment in cognitive areas that include memory consolidation, executive skills, semantic fluency and naming, and visuospatial analysis and praxis. The second most common form of dementia is vascular dementia (VaD), which often occurs with an abrupt onset and manifests itself in language and memory retrieval difficulties and executive inefficiencies, among other symptoms. Several studies, including Groves et al. (2000) and Blazer and Steffens (2009), have acknowledged the difficulty in differentiating between AD and VaD, because of the “many similarities in the clinical presentations of the two disorders” (Groves et al., 2000, p. 306). Adding to the complexity of diagnosis, other types of dementia (such as frontotemporal dementia, Lewy body dementia, Parkinson’s disease dementia, semantic dementia) may occur in combination with AD or occur alone with very similar symptoms.

While there is no proven cure for many types of dementia, a correct, timely diagnosis is of great importance. VaD, for instance, can rapidly become severe if treatment and prevention of vascular disorders (e.g., stroke, hypertension) are not promptly implemented. In addition, some types of

[2] André Picard. “Costs to soar as aging Canadians face rising tide of dementia.” *Globe and Mail*, 3 January 2010.

dementia are reversible (such as those caused by vitamin B12 deficiency, hypothyroidism, normal-pressure hydrocephalus, depression), and can be treated if correctly identified in time. Blazer and Steffens (2009) suggest that early diagnosis of AD may even make prevention possible:

> The Alzheimer's pathology likely begins many years and perhaps decades before the onset of symptoms; therefore, there is an opportunity for prevention once future advances make it possible to diagnose the disease through the use of biomarkers before symptom onset (p. 249).

Or, perhaps, through linguistic markers, as demonstrated in the Nun Study by Snowdon et al. (1996), which examined early language samples as potential predictive indicators of AD. The study involved ninety-three participants, aged 75 to 95, who were members of the School Sisters of Notre Dame. The sisters underwent an assessment of cognitive function and allowed their autobiographies, written in their early 20s, to be analyzed for idea density and grammatical complexity. The results revealed a strong and consistent association between low idea density in early written texts and low cognitive test scores assessed approximately 58 years later. Among the participants who subsequently died, AD was neuropathologically confirmed in all of those with low idea density and none of those with high idea density. The results suggest that linguistic ability in early life may strongly predict risk of poor cognitive health and AD in late life and, furthermore, demonstrate the potential of high-accuracy diagnosis of dementia that is based solely on linguistic evidence.

1.2 Language of the Elderly

In this section, we establish distinctions between possible linguistic changes found in normal aging and those found in the presence of dementia.

Several studies on language in normal aging have drawn different conclusions on which aspects of language, if any, are altered and the extent of the changes. Kemper et al. (1989) conducted a study in which language samples of three discourse genres—oral questionnaire, oral statement, and written statement—were elicited from participants belonging to two age groups: young adults and (presumably healthy) elderly adults. Significant age-related decline in the mean number of clauses per utterance and the percentage of left-branching clauses was found across all three genres of discourse when individual differences in linguistic abilities were taken into account. Since sentences containing more clauses and left-branching clauses impose greater demands on working memory than do simple sentences and right-branching clauses, these age effects were attributed to diminished working memory capacity in older adults.

On the other hand, Kemper et al. (1989) found no significant changes in the percentages of sentence fragments and lexical fillers (e.g., *well*, *yeah*, *let's see*); it is important to note, however, that the researchers excluded nonlexical fillers (e.g., *uh*, *um*) from their language samples, although this category of disfluencies could arguably reflect language processing or retrieval speed. Interestingly, while the overall syntactic complexity produced by the older age group was lower, their oral and written statements were deemed clearer and more interesting by human judges, and indeed a correlation was found between low ratings of clarity and interest and the use of complex syntactic constructions. The study concluded that, first, as remedy for their loss of memory capacity, elderly adults do not produce more sentence fragments or rely on lexical fillers, but instead restrict the syntactic complexity of their sentences. Second, this decline in syntactic complexity may also be in part due to the elder group's acquired proficiency in communication, since aspects of language, such as clarity and appeal, seem to improve with age.

Surveying studies on language of the healthy elders, Maxim and Bryan (1994) reported the following results, some of which support and some contradict the previously described findings by Kemper et al. (1989). Left-branching clauses were deemed more difficult for elderly adults to process and, in a similar cross-sectional study between two age groups, the older group was found to produce significantly fewer number of clauses per sentence and, in addition, more sentence fragments. Two other studies found a significantly higher number of disfluencies in the language of the elderly, in particular, hesitant interjections and fillers (presumably of either type, lexical or nonlexical). This increase was suggested to indicate slower language processing time, or vocabulary retrieval problems while language complexity and organization were arguably well preserved. This final point, regarding the preservation of complexity, was contradicted by another hypothesis reported by Maxim and Bryan (1994): that a reversal of children's language acquisition process at the grammatical and semantic levels occurs with advancing age. Evidence of this linguistic regression in both normal aging and AD was demonstrated in several studies, including that by Kemper et al. (2001).

In their research, Kemper et al. (2001) conducted a longitudinal study following linguistic changes in healthy elders and dementia patients. Language samples, vocabulary scores and digit span scores were measured annually for the former group over a period of 7–15 years, and semi-annually for the latter group for up to 2.5 years. Two measures were used to analyze the language samples: P-Density, to estimate the density of propositional content (i.e., how much information is conveyed relative to the number of words), and D-Level, which models stages of children's language development, to evaluate the grammatical complexity of each complete sentence. Both measures revealed a cubic function[3] of decline in the writings produced by the healthy group, modeling three periods: one of relative stability, followed by rapid decline in the participants' mid-70s, and finally a period of more gradual decline. The writing samples by the group with dementia, although produced over a much shorter time span, displayed an accelerated rate of decline in both measures, which coincided with the onset of dementia and was better modeled as a function of time rather than age. While propositional content exhibited a linear pattern of decline, grammatical complexity followed a cubic model similar to that of the healthy group. However, some dementia patients registered slight improvements in their final assessments, which were deemed a possible effect of pharmaceutical intervention. The study concluded that dementia may precipitate linguistic decline, along with the deterioration of cognitive abilities; thus the general pattern of change in dementia patients would exhibit a relatively steeper drop, beginning at the disease onset.

In another comparative study, Bird et al. (2000) used the Cookie Theft picture description task[4] to elicit oral narrative samples from three semantic dementia patients and twenty normal controls. As their disease progressed, the former group exhibited increasing difficulty in word retrieval,

[3] A cubic function is a polynomial of degree three. An example of a cubic model is shown in Figure 18b (p. 69).

[4] In this task, participants are asked to describe a drawing that depicts a domestic setting, in which a female adult is seen washing the dishes, with water overflowing from the sink, oblivious to a boy behind her who is falling off a stool while handing cookies from a jar to a young girl. Preciseness of vocabulary and coverage of detail are among the variables often measured by studies that employ this procedure.

and words belonging to the lower frequency bands progressively disappeared from the patients' vocabularies. In the early to moderate stages of the disease, a prominent deficit was observable in nouns while verbs, generally higher in frequency ranks than nouns, were relatively spared. When the disease became severe, however, verbs began to be affected, leading to a reverse imageability effect, since very high frequency verbs—including *be*, *come*, *do*, *go*, *have*, *think*—are generally of lower imageability. The study proposed that the vocabulary of a semantic dementia patient would gradually shrink until it contains only the most frequently occurring words which, comprising mainly verbs and very few nouns, tend to be very general.

These results are consistent with other studies into language of AD patients reported by Maxim and Bryan (1994). Among the most noticeable deficits at the semantic level, as well as the earliest symptoms of the disease, is a reduction of available vocabulary, leading to increasing difficulty in word finding. When presented with an item (e.g., an image of a cat) along with a list of words containing the name of this item, a participant diagnosed with possible AD was more prone to error in selecting the correct name when the list contains a semantically related item (e.g., *dog*) than when the list items are semantically distant (e.g., *cat*, *kettle*, *boat*, *pencil*). This was suggested to indicate "the gradual loss of semantic features for each item, with more specific semantic features lost before more general semantic features" (p. 182). In other words, although able to recognize that *cat* and *dog* both belonged to the more general semantic category *animal*, the patient had trouble picking out the correct term of higher specificity within the same category.

In addition, increased lexical repetition is also a well-documented phenomenon in the language of dementia patients (Nicholas et al., 1985; Smith et al., 1989; Holm et al., 1994). Ideas from previous utterances are often reiterated in the same words, phrases, or even short sentences, either as perseverations or as "markers when other lexical items are not available" (Maxim and Bryan, 1994, p. 183), the latter being a direct result of the marked reduction in active vocabulary.

At the syntactic level, contrary to the popular belief that grammar is spared in AD, Maxim and Bryan (1994) reported the observations that patients often had trouble with "understanding and constructing complex grammatical structures," and that "the process of reduction to more simple semantic and grammatical forms also took place," partly because of the patients' "difficulty with retention of grammatical complexity" (p. 185). Bates et al. (1995) suggested that, while syntax may appear intact and AD patients may still be able to produce well-formed sentences, deficits may emerge under constraints. The researchers compared the use of passive voice among three groups of participants: one consisting of 16 patients diagnosed with probable AD, a control group of 25 healthy elders, and a young control group of 11 undergraduate students. The participants were asked to describe the actions or events shown in several animated clips, first without constraints, and then with focus on the objects of the actions (through the use of prompts such as "Tell me about the ____"). The latter of these procedures provided a natural context for passive voice, and indeed, all three groups produced almost all of their passive forms in response to the probes; however, there were large quantitative differences among the groups. In their descriptions of 24 clips, the young and elderly control groups produced 14.81 and 11.6 passives, respectively; the AD group, on the other hand, produced an average of 5.31 passives, and one third of the group failed to produce any passive form in their responses. Differences were also found in the types of passive produced. The AD group used more agentless passives (e.g., *John was fired*, or *John got fired*) than either of the control groups: only 84.3% of their passive forms contained a *by*-phrase (e.g., *John was/got fired by his boss*), compared to 90.6% and 97.2% for the elderly and young control groups, respectively. The AD group also relied heavily on the *get* form of passive, which accounted for 64.1% of all passive forms, in contrast with 29.1% for the elderly controls and 21% for the young controls. The study concluded that syntactic production deficits do occur in AD, leading to differences in both the number and the types of passivisation compared to age-matched controls, while only quantitative differences existed between the elderly and the young controls, which suggested that healthy elders retained access to the same range of syntactic forms, but utilized the alternative forms less often.

In summary, heterogeneity is expected in the linguistic changes among individuals in both normal aging and dementia. While different studies have offered different theories regarding the linguistic components that undergo change, the consensus is that any decline that may occur in normal aging is accelerated in the presence of dementia. The distinguishing feature between a disease-related linguistic deficit and the natural decline associated with advancing age, then, is the rate of change, which is more gradual and less severe in healthily aging adults. In the case of dementia, deficits in lexical features may be more prominent than in syntactic ones, since a core of linguistic ability is possibly spared until the later stages of the disease progression.

Table 1 summarizes the reported linguistic changes in normal aging versus those in dementia. The items presented in parentheses are not explicitly stated in the previous studies; for instance, Maxim and Bryan (1994) reported the loss of semantic features observed in AD without comparison to normal controls. Several of the lexical linguistic markers in normal aging indicate possible change in either direction, the reason being that, according to several studies reported by Kemper et al. (2001), one's vocabulary increases throughout the middle adult years but may decrease in late adulthood even without the presence of cognitive disease. Whether this decrease has begun determines the direction of change in lexical markers L1, L2, and L3. A smaller active vocabulary may lead to a higher rate of content word repetitions, since fewer words are available, and a lower degree of word specificity, since one is more likely to resort to common, general words and phrases.

Table 1: Patterns of linguistic changes expected in normal aging and dementia

LINGUISTIC MARKER	NORMAL AGING	DEMENTIA
Lexical:		
(L1) Vocabulary size:	gradual increase, possible slight decrease in later years	sharp decrease
(L2) Lexical repetition:	(possible slight decrease/increase)	pronounced increase
(L3) Word specificity:	(possible slight increase/decrease)	pronounced decrease
(L4) Word class deficit:	insignificant change	pronounced deficit in nouns; possible compensation in verbs
(L5) Fillers:	possible slight increase	(pronounced increase)
Syntactic:		
(S1) Overall complexity:	no change or gradual decline, possible rapid decline around mid-70s	sharp decline
(S2) Use of passive:	possible slight decrease	pronounced decrease

1.3 Hypothesis

We hypothesized that most, if not all, of the patterns given in Table 1 are present in the writings of healthy elders and dementia patients. More specifically, with respect to our selected authors, we hypothesized the following:

- P. D. James's novels will exhibit the linguistic patterns of normal aging.
- Iris Murdoch's novels will exhibit the linguistic patterns of dementia patients.
- Agatha Christie's novels will resemble the patterns found in Murdoch's novels.

1.4 Organization

After a survey of similar works in section 2, we provide a more detailed introduction of the selected authors and the basis for this selection in section 3. The techniques employed in our analysis are described in section 4, the results presented and discussed in section 5. Section 6 closes with a summary of our findings, the limitations of our approach, and suggestions for future development.

2 Related Works

In recent years, a few longitudinal studies have been conducted with focus on individual writers, in order to examine their patterns of linguistic changes over time. Williams et al. (2003) analyzed fifty-seven letters written by the seventeenth-century monarch, King James VI/I, within the last twenty years of his life to assess whether the linguistic cues in these letters reflected normal aging, AD or VaD. The researchers relied on type/token ratio (TTR) to analyze semantic complexity, and computed the mean length per utterance (MLU), the mean number of clauses per utterance (MCU) and the average D-Level score[1] to measure grammatical complexity of the texts. The results revealed a quadratic pattern[2] of decline in syntactic complexity, as reflected by MCU (but not MLU or D-Level), and increased diversity of vocabulary, as reflected by TTR, beginning in James's early fifties. This prompted the suggestion that James was relying on semantic functions to compensate for the decline in syntax. Backed by medical records and autopsy results, the researchers observed that James may have suffered from chronic hypertension—a condition that may be an antecedent to VaD (Posner et al., 2002). The study did not produce a conclusive diagnosis, because of "the unknown applicability of modern linguistic analysis to Elizabethan writing style" (p. P44), as well as the difference in health and life span in the seventeenth century, which may have resulted in a different pattern of cognitive change, as the researchers cautioned.

In a similar case study on a contemporary subject, Garrard et al. (2005) examined works by the late English author Iris Murdoch, whose diagnosis of AD a few years before her death was later confirmed post mortem. Murdoch's last novel *Jackson's Dilemma*, published shortly before her diagnosis, is widely believed by researchers, literary critics, and readers to contain indicators of the acclaimed author's declining cognitive health. Along with this novel, Garrard et al. (2005) sampled two of Murdoch's earlier works: her first published novel *Under the Net*, and one written at the height of her career, *The Sea, the Sea*. The researchers conducted a thorough study into Murdoch's biographies and approach to writing, and presented neuropsychological test results

[1] The D-Level measure was incorrectly listed under "Semantic Complexity" (p. P43) despite being introduced as "an indicator of grammatical complexity" (p. P42).

[2] A quadratic function is a polynomial of degree two. An example of this model is shown in Figure 18b (p. 69).

and brain scan images—a unique and valuable contribution to the studies of dementia effects on language. However, the linguistic analysis conducted suffers from problems in methodology.

The first problem lies in the data used for analysis. The complete texts of *Under the Net* and *Jackson's Dilemma* were digitized; for undocumented reasons, only the first 100 pages of *The Sea, the Sea* (approximately one fifth of the novel) underwent the same process. This affects the reliability of the structural analysis, which compared the novels in terms of the total length in words, the number of chapters, the number of characters, and the narrative/dialogue ratio. Apart from the number of chapters, these measures were computed for *The Sea, the Sea* based on estimates "projected from a word count of 41,817 words in the first 100 pages" (p. 254); for instance, this number was multiplied by 5 to approximate the word count of the entire novel. This projection relied on the questionable assumption that the remaining 400 pages of the novel were identical in structure to the first 100 pages (indeed, the projected word count was off by nearly 7000 word tokens).

Even if this problem with the data could be overlooked, we question the relevance of these measures, since little implication on the cognitive health of the author could be drawn from facts such as "the first book is subdivided into more chapters than either of the two later works, while the middle work is far and away the longest of the three" (p. 254). The variables measured in this structural analysis depend heavily on the topics, genres, settings, perspectives, plot development and other characteristics, which may vary among works by the same author regardless of age or cognitive health. For instance, a story written in the first-person may differ vastly in terms of narrative/dialogue ratio compared to one written in third-person. (Coincidentally, *Under the Net* and *The Sea, the Sea* were told from the first-person perspective, while *Jackson's Dilemma* was a story in third-person. The reported narrative/dialogue ratios for these novels were 0.18, 0.13[3] and 0.26, respectively.) Furthermore, differences in structure may stem from stylistic choices or experiments, which may have been a factor considering Murdoch's highly praised talent for transforming her style and narrating convincingly in the voices of her many and diverse characters.

From the available texts, the researchers compiled two word lists for each novel: a list contain-

[3] Based on the first 100 pages of *The Sea, the Sea*.

ing all word tokens and an incomplete one, generated by randomly selecting 100 words five times, and thus containing 500 word tokens or, discounting duplicates, 352 to 379 unique word types. (By our computation, each of the three novels contains a total of 5045 to 9076 unique word types.) These abridged word lists, arguably unrepresentative samples of the novels owing to their small sizes and the randomness of their selection process, were used as data for most of the measures computed by hand, namely, average word length, average word frequency, and grammatical class proportions. The results of these measures are therefore not statistically reliable. One puzzling fact about the chosen methodology is that at least one of these measures could have been automated on the full word lists without compromising accuracy, in particular, the average word length measure. In addition, the grammatical class of each word was not determined by part-of-speech tagging, but rather, "the more typical reading" out of four categories (noun, verb, descriptor and function word) was selected, regardless of context (p. 253). The accuracy of this approximation is uncertain, since relatively few English words belong to only one word class. In the following passage from *The Sea, the Sea*, the underlined words belong to more than one category,[4] and the doubly underlined are words whose parts of speech in this context are different from their more typical readings:

> We are in the north, and the bright sunshine cannot penetrate the sea. Where the gentle water taps the rocks there is still a surface skin of colour. The cloudless sky is very pale at the indigo horizon which it lightly pencils in with silver. Its blue gains towards the zenith and vibrates there. But the sky looks cold, even the sun looks cold.

For the remaining measures performed by hand, Garrard et al. (2005) used a different sample set. The first ten sentences from the first, middle and final chapters of each book were extracted as data for two measures of syntactic complexity, the mean number of words per sentence and the mean number of clauses per sentence. The rationale for this choice—that the segments were "similarly sized samples from equivalent points in the three books"—is unsound; as the researchers themselves pointed out, the results of these measures "would have been influenced not only by the book's overall syntactic complexity, but also by the local thematic context" (p. 255). Furthermore,

[4] Based on the non-obsolete, non-archaic entries in the Oxford English Dictionary, second edition, 1989.

a chapter may begin with a narrative, a dialogue, a letter, etc., regardless of its location in the novel; these types tend to have different levels of syntactic complexity. Considering the first ten sentences of the first chapter, *Under the Net* opens with a somewhat colloquial, first-person narration in the voice of the main character; *The Sea, the Sea* quotes the opening paragraph of the main character's memoir, written in quasi-poetic language describing his view of the sea; *Jackson's Dilemma*, on the other hand, starts with a realistic description of a character from a third-person point of view. Given the differences in perspective, tone and thematic context, it is difficult to determine whether the reported syntactic differences were due to conscious stylistic choices or inevitable cognitive decline. To account for this problem, the researchers automated an approximation of the mean number of words per sentence over the entire texts (or the first 100 pages, in the case of *The Sea, the Sea*). This was achieved by dividing the number of words by the number of sentence-ending markers (periods, exclamation marks and question marks). One complication is that these punctuation marks do not necessarily indicate the end of a sentence, for instance, periods in abbreviations or initials, and exclamation marks or question marks in direct speech (e.g., from Murdoch's *Under the Net*: "I shouted 'Hey!' and Finn came slowly on."). Whether these cases were excluded from the count of sentence-ending markers is unknown.

The remaining measures included a variation of type/token ratio, which computes the number of unique word types at every 10,000 word-token interval. This measure revealed an impoverishment in vocabulary in the first 40,000 tokens[5] of Murdoch's last novel, *Jackson's Dilemma*, relative to similar-sized portions of the two earlier novels, and a slower rate of new word-type accretion in *Jackson's Dilemma* compared to *Under the Net*. The final measure, termed *auto-collocations*, computes the proportion of times the ten most common words were repeated within a space of four subsequent words.[6] The analysis showed that the most common repetitions were function words (e.g., *the*, *a*, *and*, *of*); this finding, however, led to no definite conclusion. Overall, the study contended that, while few disparities were found in the structure and the syntax, marked and consistent variations existed in the lexical analysis of small samples randomly drawn from the novels.

[5] This length restriction was due to the size of the incomplete text of *The Sea, the Sea*.

[6] It was unclear over what total this proportion was computed.

3 Materials

As mentioned previously, we chose the writings of Iris Murdoch, Agatha Christie, and P. D. James as data for textual analysis. We collected a total of fifty-one novels, spanning a minimum of four decades over each author's life, to represent the authors' writing careers. A complete listing of the novels, along with their publication dates and the estimated ages of the authors at the time of composition, is given in Appendix A. In this section, we give a brief introduction of the authors, why we selected their writings, and how the data was obtained.

3.1 Authors

Iris Murdoch (1919–1999) was an English novelist, philosopher, playwright and poet, best known for her critically acclaimed novels covering a wide range of topics, such as moral dilemmas, personal struggles, and sexual identities. Murdoch's novels have received many prestigious awards in literature, including the Booker Prize[1] in 1978 for *The Sea, the Sea* and the James Tait Black Memorial Prize[2] in 1973 for *The Black Prince*. Her first published work, *Under the Net*, was listed among the 100 best English-language novels of the twentieth century by the editorial board of the American Modern Library,[3] along with the likes of James Joyce's *Ulysses*, Aldous Huxley's *Brave New World*, George Orwell's *1984*, and Ernest Hemingway's *A Farewell to Arms*.

Towards the end of her career, however, Murdoch began to encounter difficulties in writing that she first attributed to a period of writer's block, according to her husband, John Bayley, in his 1999 memoir, *Elegy for Iris* (see Garrard et al., 2005). At the time, Murdoch was working on *Jackson's Dilemma*, which ultimately became her last novel. Published in 1995, the book received mixed reviews that were often less than favourable—a sharp contrast to her previous works. While some critics thought *Jackson's Dilemma* showcased Murdoch "at the height of her powers,"[4]

[1] The Man Booker Prize. *Prize archive*. http://www.themanbookerprize.com/prize/archive

[2] The University of Edinburgh. *The James Tait Black Memorial Prizes*. http://www.englit.ed.ac.uk/jtbwins.htm

[3] Modern Library. *100 Best Novels*. http://www.randomhouse.com/modernlibrary/100bestnovels.html

[4] Geoffrey Heptonstall. *Contemporary Review*, 1 January 1996.

"afire" with a fast-paced story which was "almost all plot, no decoration, no reflection,"[5] others found the novel "hard to digest,"[6] the narrative moving "with scant explanation,"[7] the plot "a thin trickle of unconvincing incidents"[8] and the writing "a mess,"[9] "bad beyond belief"[8] with phrases such as *then suddenly* "appearing three times in a single paragraph."[9]

Diagnosed with AD shortly after this final publication, Murdoch passed away in 1999, donating her brain to science to help advance research into the disease. Her diagnosis was confirmed by a postmortem examination. Long suspected to contain signs of the author's cognitive decline during the development of her disease, *Jackson's Dilemma* contributes significant data that enables further research into effects of AD on language and writing. In addition to this novel, we have collected and digitized another nineteen of Murdoch's twenty-six novels, published between ages 35 and 76 [M=52.7]. The linguistic patterns observed in this series of novels served as the patterns of changes in the written language of dementia patients.

Agatha Christie (1890–1976) was a renowned English novelist, playwright and short-story writer, whose prolific 53-year career produced an impressive collection of 90 novels, 15 plays, and 147 short stories. Her crime fictions, especially those featuring detectives Hercule Poirot and Miss Marple, for which Christie was most famous, earned her the undisputed title "the Queen of Crime." To date, she remains the best-selling fiction author of all time, according to the Guinness Book of Records, outsold only by the Bible and William Shakespeare.[10] Christie is recognized by UNESCO's Index Translationum statistics as the most translated individual author, with her books translated into at least 56 different languages.[10]

Known for her consistent, "unassuming, and colloquial but not slangy style" (Lancashire, Forthcoming 2010) Christie laid out in each of her crime fictions a mystery, sprinkled with clues and diversions to keep the readers enthralled, often leading them down the wrong track, only to

[5] Carey Harrison. *San Francisco Chronicle Book Review*, 24 December 1995.

[6] Valerie Miner. *Nation*, 8 January 1996.

[7] Kate Kelloway. *Observer*, 1 October 1995.

[8] Merle Rubin. *Wall Street Journal*, 12 January 1996.

[9] Brad Leithauser. *New York Times Book Review*, 7 January 1996.

[10] Wikipedia. *Agatha Christie*. http://en.wikipedia.org/wiki/Agatha_Christie

have her shrewd detective reveal the real fiend(s) at the close of the story and solve the mystery. Her final novels, however, failed to achieve these qualities that had become Christie's trademarks. *Postern of Fate*, published in 1973 when Christie was 82, was considered by critics "a contrived affair that creeps from dullness to boredom,"[11] its plot "total chaos" and its clues "total confusion."[12] Sage (1999) found this last novel, along with *Elephants Can Remember* (1972), "execrable" and that Christie as a writer had "[lost] her grip altogether" (p. 132).

Although Christie was never formally diagnosed, the unusually harsh criticisms of her later works amount to compelling evidence of a decline in the quality of her writings toward the end of her life and career. This fact was confirmed in the biography of Christie by Janet Morgan, who was granted access by Christie's daughter to the author's letters, manuscripts, and diaries, among other private documents.[13] On *Postern of Fate*, Morgan (1984) observed that Christie's "powers really declined": she "found it harder than ever to concentrate," and according to her husband, Max Mallowan, finishing this last book "nearly killed her" (pp. 370–371). Christie herself felt "uneasy" about the result; she asked Mr. Edmund Cork, her agent, for "a candid opinion," to which he suggested she have some help with the novel (p. 371). Christie eventually did—perhaps for the first time in her career—seek editing help from her husband and their secretary, Mrs. Daphne Honeybone, who "tidied it up" (p. 371).

These facts, in and of themselves, are not evidence that Agatha Christie was plagued by cognitive diseases, although they clearly point in that direction, considering the dramatic decline in writing prowess of an author of her stature. To assess the possibility of dementia, we collected sixteen of Christie's novels written between ages 28 and 82 [M=59.0] and compared their patterns of linguistic changes to those of Murdoch's novels. In addition, to rule out the possibility that the decline was associated with normal aging, these patterns were contrasted against those found in the novels of a third author, P. D. James.

[11] Newgate Callendar. *New York Times Book Review*, 1973.

[12] Harry C. Veit. *Library Journal*, 1 January 1974.

[13] Review of Janet Morgan's *Agatha Christie: A Biography*. Kirkus Reviews.

Phyllis Dorothy James White (born 1920), commonly known as P. D. James, is an English novelist. Crowned the Queen of Crime in contemporary fiction following Agatha Christie's reign, James has been praised as "the most literary of crime writers,"[14] with her "clear, stylish prose"[15] and "nuanced portrayal" of characters in "stories that were novels first, mysteries second."[16] Having written twenty-one books since her debut in 1962, James is best known for her novel series featuring her iconic creation, "the most misspelt senior policeman in crime fiction,"[14] Adam Dalgliesh. Credited for "[having] elevated English detective fiction far beyond the diverting puzzles typical of the genre novelists of an earlier generation,"[17] James has received many literary honours, including the 1999 Grand Master Award from Mystery Writers of America.[18] She was the third writer to be inducted into the International Crime Writing Hall of Fame in 2008,[18] along with Arthur Conan Doyle and Agatha Christie.

At 89 years of age and still at the height of her career, James appears to be a remarkable example of a healthily aging elder. "I was extraordinarily lucky with health," she said in an interview, "I really didn't feel particularly old. We don't grow gradually into old age. Throughout our lives, we're on a plateau and then suddenly, whoosh! We're five years older, and then we're on a plateau again."[19] The *whoosh* moment, to which James was referring, happened in 2007 when, after a hip replacement, she suffered a heart failure. Yet her health condition did not stop her from writing. Her latest novel *The Private Patient*, whose theme is set in a private clinic for plastic surgery, was in fact inspired by James's stay in an Oxford convalescent hospital during her recovery. "I've never known the last part of a book go so easily," the prolific author recalled.[19] The novel was published in 2008 to a generally favourable reception. While its plot may not be "up to this author's diabolical best,"[20] "the characterisation, the accretion of detail, the overarching humanity is as impressive as ever."[21] Written in James's late 80s, *The Private Patient* "shows no signs of author fatigue,"[14]

[14] Marcel Berlins. *Times*, 30 August 2008.
[15] Simon Akam. *New Statesman*, 21 August 2008.
[16] Michael Norman. *Plain Dealer*, 29 November 2008.
[17] *Times*, 10 September 2008.
[18] The Official Website of P. D. James. http://www.randomhouse.com/features/pdjames/abouttheauthor.html
[19] "PD James: Heroine with a taste for life." *The Independent*, 29 August 2008.
[20] Janet Maslin. *New York Times*, 19 November 2008.
[21] David Robson. *Telegraph*, 7 September 2008.

"propelled, as always, with James's eloquent way with words,"[22] "consummate skill and delicious irony."[23] Continuing to produce critically acclaimed works, in 2009, James examined the craft that she has mastered—and helped revolutionize in a fifty-year writing career—in her most recent book, *Talking about Detective Fiction*. In an interview with BBC director general Mark Thompson on the Today program in December 2009,[24] James "skewered him with the sheer force of her brain and her indignation."[25] James, "as sharp as a razor" and with "a mind like a steel trap,"[25] was widely praised for this interview, in which she reduced the director general to "a stuttering wreck"[26] and echoed the sentiments of many BBC listeners and viewers.[27]

Choosing James's writings as a linguistic model for written language in normal aging, we collected fifteen of her novels, published between ages 42 and 82 [M=63.9]. The patterns observed will be contrasted with those of Murdoch's and Christie's novels in a comparative analysis of lexical and syntactic features.

[22] Louise France. *Observer*, 7 September 2008.

[23] Nicholas A. Basbanes. *Los Angeles Times*, 22 November 2008.

[24] "When PD met DG Mark Thompson." *Today*. BBC Radio 4, 31 December 2009. http://news.bbc.co.uk/today/hi/today/newsid_8435000/8435731.stm

[25] Sarah Thompson. "PD James and the BBC: Here at last was someone saying what so many people feel." *Telegraph*, 1 January 2010.

[26] Sam Greenhill. "PD James handbags BBC chief on sky-high salaries." *Daily Mail*, 1 January 2010.

[27] Philip Johnston. "BBC pay, bureaucracy and ageism: PD James speaks for the nation." *Telegraph*, 31 December 2009.

3.2 Data

Sources:

Two of Christie’s novels, *The Mysterious Affair at Styles* (1920) and *Secret Adversary* (1922), were obtained from Project Gutenberg, a website that digitizes materials that are not, or no longer, copyrighted and makes them available online for public use. The remaining forty-nine novels[28] by Murdoch, Christie, and James were scanned and digitized with commercial optical character recognition (OCR) software.[29] Lexical and punctuation errors made by the OCR were corrected manually, and then semi-automatically using an interactive program designed to identify common patterns of errors. Examples of common OCR errors and their corrections are given in Table 2.

OCR Error	Correct text
arid	and
die	the
gende	gentle
comer	corner
Til	I’ll
AH	All
6	‘
9	’
.’	*?*

Table 2: Common OCR errors

Appendix A lists the fifty-one novels that were analyzed, with publication years and estimated ages of the authors at the time when the novels were written. Apart from Christie’s *Curtain*, which she wrote during World War II and then laid aside, unpublished, as a source of income for her family in the future (Lancashire, Forthcoming 2010), the remaining novels are assumed, as there is no evidence to the contrary, to have been written relatively close to the time of publication.

[28] Of these, fourteen Agatha Christie novels were digitized by Dr. Ian Lancashire and his student, Mr. Tim Harrison, who kindly provided us with the texts for our analysis.

[29] OmniPage Professional 15.0 for the fourteen Christie novels, and ABBYY FineReader 9.0 Professional Edition for the novels by Murdoch and James.

Also included in Appendix A is a unique identifier for each novel, containing the first letter of the corresponding author's last name and a number, based on the order of composition. These identifiers are henceforth used to indicate the source texts from which example sentences are drawn. For instance, [M1] signifies that the source text is Murdoch's first novel, *Under the Net*.

Impact of Subsequent Revisions and Reference Sources:

An inevitable process behind any published work is the editing that follows the first draft, either by the authors or by their editors. The issue with which we are concerned is the amount of editing performed on the selected novels, since this process potentially affects the lexical and, arguably to a lesser degree, syntactic properties of the texts. While this issue cannot be resolved with certainty, public information regarding the writing processes of the authors addresses it to some extent.

Garrard et al. (2005), drawing on biographies of Murdoch and the memoir by her husband, emphasized that the author "regarded the manuscripts that she sent to her publishers as representing her work in its finished form, and resisted any suggestions of alterations to the text" (p. 252). In addition, Murdoch wrote her manuscripts in longhand after months of working out the plot, without using a typewriter or word processor, and apparently neither "agonized over choice of words, indulged in repeated revisions of passages, [nor] made extensive use of a dictionary or thesaurus" (Garrard et al., 2005, p. 252).

P. D. James's approach to writing also involves several months of sketching the plot in a notebook, writing out-of-order sequences of the novel by hand, putting the book together and dictating it to a secretary who types into a computer.[30] Little is known about the extent of revisions, editorial interference, or reliance on reference sources in James's final products, aside from the fact that a dictionary and a thesaurus are among the items always on her desk.[31] However, given James's previous comment that she found the writing process of her latest novel easy, there is no reason to suspect that she relied on reference sources in her later novels any more than she did in her earlier ones; thus a comparative analysis on her novels remains valid.

[30] The Official Website of P. D. James. http://www.randomhouse.com/features/pdjames/faq.html

[31] The Official Website of P. D. James. http://www.randomhouse.com/features/pdjames/abouttheauthor.html

In contrast, Agatha Christie's writing process changed over time. She wrote her earlier novels in longhand and then typed them on a typewriter; then, for a brief period, she hired a secretary to type according to her dictation (Lancashire, Forthcoming 2010). After an accident in 1952 in which she broke her wrist, Christie started using a dictaphone and, while others thought the device improved her writing (Lancashire, Forthcoming 2010), she found the subsequent revisions, required to remove repetitions made during the recording, "irritating," and wrote in her autobiography that dictation "destroys the smooth flow which one gets otherwise" (Christie, 1977, p. 348). These comments by the author herself imply that very little editing was done to the previous works, to preserve the smooth flow of the original draft—indeed, Christie's notes for a number of her novels "are almost identical to the finished article" (Thompson, 2007, p. 369); how extensively revised her post-1952 novels were is unknown. Her final novel, *Postern of Fate*, is known to have involved editing help from family friends at the request of her agent (Morgan, 1984, p. 371). These changes in Christie's writing process are annotated in the list of her novels used in our analysis (see Appendix A) and should be taken into consideration when interpreting the analysis results. With respect to reference sources, Christie's novels were based largely on her own knowledge and experiences, often inspired by her travels to different countries and trips to archaeological sites with her husband Max Mallowan.[32] One exception is *Passenger to Frankfurt: An Extravaganza* (1970). In writing this novel, Christie did extensive research into political literature with the help of her publisher, and the book itself, being a thriller, belongs to a different genre compared to the other novels in our dataset; as a result, *Frankfurt* was enriched with a vocabulary beyond that of her own (Lancashire, Forthcoming 2010). To account for this, we considered this novel an outlier and excluded it from the dataset for measures that assess natural changes in vocabulary.

To the best of our knowledge, none of the remaining novels in our datasets departs from the usual writing methodology of its author or involves an extraordinary amount of research to such an extent that it should be deemed an outlier.

[32] The Official Website of Agatha Christie. http://www.agathachristie.com/about-christie/travel-and-archeology/

4 Methods

4.1 Lexical Measures

We now describe the measures developed to assess the lexical patterns of change given in Table 1 (p. 15). Our implementation employs NLTK WordNet's *morphy* method (Bird et al., 2009) for lemmatization, and the Charniak parser (Charniak, 2006) for part-of-speech tagging; section 4.3 contains a more detailed discussion of the implementation. In analyzing each dataset, we must take into account superficial differences across texts and whether these variables could affect the results. Measures that are sensitive to length (e.g., counting the number of occurrences of some phenomenon) must either report results as percentages (rather than absolute counts), or have a cut-off threshold that is at most the length of the shortest text, in which case only a portion of each text, from the beginning up to this threshold, is considered. Each text in our dataset contains at least 55,000 tokens, except for Murdoch's *The Italian Girl*, with a word count of 48,448. Thus, in addition to the lexical outlier *Passenger to Frankfurt*, we also excluded *The Italian Girl* when computing measures that are normalized in length, to avoid lowering the length threshold.

L1: Vocabulary Size

Change in vocabulary size is assessed by the type/token ratio measure, calculated by dividing the number of unique *lemmatized* word types by the total number of word tokens. While it is not immediately obvious that this measure is sensitive to text lengths, we generally do not expect a 60,000-word novel to have twice as many unique word types as, for instance, a 30,000-word novella, since the number of word types does not grow in proportion to the number of word tokens (the second half of the 60,000-word novel is bound to "reuse" a large number of the word types found in the first half). Thus type/token ratio for each text is computed only up to the 55,000th token. To avoid this length restriction, we measured the word-type introduction rate, in an approach similar to that of Garrard et al. (2005). This measure reports the number of unique lemmatized types computed at every 10,000th token, to compare the vocabularies of equal-sized portions of the texts, and also the rates of growth of word types with respect to tokens.

L2: Lexical Repetition

As reported in section 2, Garrard et al. (2005) also computed the proportion of times the ten most common words in each text were repeated within a space of four subsequent words. Since the most common repetitions are inevitably function words (e.g., *the*, *a*, *and*, *of*), this measure was classified as a syntactic analysis technique; however, little conclusion can be drawn from its results in terms of syntactic complexity. The following sentence, for instance, contains three repetitions within four subsequent words, a fact that reveals little about the writer's syntactic level:

> Unless one is very talented indeed there is no resting place between the naive and the ironic; and the nemesis of irony is absurdity. [M15]

When modified to only consider repetitions of only content words (i.e., those tagged as noun, content verb, adjective or adverb) within ten subsequent lemmatized content words, the measure becomes a lexical analysis of an author's tendency to repeat words within close distance. While an author sometimes uses deliberate repetition for effect, an increasing rate of repetition in the long term may indicate a reduced vocabulary or word retrieval difficulties. Implemented with a length threshold, the measure reports the absolute count of repetitions within the first 55,000 word tokens of each text and, in addition, the percentage of such repetitions over the total number of content word tokens within this portion. For example, the following passage from Christie's *Postern of Fate* contains 48 content word tokens, 32 lemmatized content word types, and 7 close-distance lexical repetitions (14.6%). The underlined words are those tagged as content words; the doubly underlined ones either are repeated or are the repetitions themselves. Because lemmatization is used, the pairs *look–looked* and *was–were* are also considered repetitions.

> She got near the door. She stopped suddenly, then walked on. It looked as though something like a bundle of clothes was lying near the door. Something they'd pulled out of Mathilde and not thought to look at, Tuppence wondered. She quickened her pace, almost running. When she got near the door she stopped suddenly. It was not a bundle of old clothes. The clothes were old enough, and so was the body that wore them. Tuppence bent over and then stood up again, steadied herself with a hand on the door. [C16]

L3: Word specificity

By specificity with respect to nouns, we refer to the relative rank of a word, determined by the size of the entity set which this word represents. Specificity of one word is determined in relation to another: the word *X* is deemed more specific than the word *Y* if *X* denotes a smaller set of entities than *Y*. *Dog*, for instance, is more specific than *animal*, but less so compared to *poodle*. Specificity is contrasted with generality rather than abstractness; indeed, one abstract entity may be more specific than another, such as *sadness*, *hunger*, and *exhaustion* compared to *feeling*, *condition* and *state*. However, when words belonging to different categories are compared, such as *counterfactuality* and *chihuahua*, their relative specificity ranks are more difficult to determine.

We rely on tree depths in WordNet to approximate the specificity ranks of nouns over each entire novel. The rationale of this approximation is that WordNet (version 3.0) organizes nouns into a hierarchy of hypernym–hyponym relationships, with a single root, *entity*, at the top level. In this hierarchy, any given word—or, more precisely, word sense—subsumes all the word senses in its sub-branches, to which it is a hypernym (either direct or inherited). Thus in theory, a greater WordNet depth implies higher specificity. Using this assumption, our measure computes the depths of all noun tokens in each text and reports the average over the entire text length.

This procedure cannot be applied to other content word classes, because their WordNet structures are not suitable for our purpose (see Appendix B.2 for a detailed discussion). As far as we are aware, there are no specificity ranking systems for adjectives and adverbs. With respect to verbs, our notion of specificity is comparable to imageability. For instance, *stride* is considered more specific than *walk*, since the former conveys the manner in which the action is carried out. Thus, to estimate verb specificity, we computed the proportion of high-frequency, low-imageability verb tokens in each text; a higher percentage indicates more reliance on generic verbs and, consequently, a lower overall specificity rank. We used the list of 14 verbs of high frequency observed in the writing samples of semantic dementia patients (Bird et al., 2000), and extended the list with 21 more verbs of relatively low specificity which may be common in narratives:

be, *come*, *do*, *get*, *give*, *go*, *have*, *know*, *look*, *make*, *see*, *tell*, *think*, *want* (Bird et al., 2000);

ask, *feel*, *find*, *forget*, *happen*, *hear*, *like*, *live*, *mean*, *meet*, *put*, *remember*, *run*, *say*, *seem*, *speak*, *suppose*, *take*, *use*, *walk*, *wonder*.

Our algorithm records the number of occurrences of the base and conjugated forms of these verbs within the first 55,000 words, and reports their percentages over the number of verb tokens.

L4: Word class deficit

After performing part-of-speech tagging on each text, we computed the proportions of each word class over the entire length of the text, both in terms of word tokens (to look for signs of deficit in or reliance on individual classes) and word types (to measure vocabulary size of open classes). Word classes of interest and the corresponding tags used by the Charniak parser are:

- common noun: NN (singular), NNS (plural);
- content verb: VB (base form), VBP (non-third-person-singular present), VPZ (third-person singular present), VBD (past tense), VBG (present participle), VBN (past participle);
- adjective: tagged as JJ, JJR (comparative form), JJS (superlative form);
- adverb: tagged as RB, RBR (comparative form), RBS (superlative form);
- pronoun: PRP (personal pronoun), PRP$ (possessive pronoun).

L5: Fillers

For this measure, we computed the proportion of interjections and fillers (those tagged as *UH* by the Charniak parser). The parser, operating on a per-word basis, only identifies single-word interjections and fillers (e.g., *well*, *yeah*, *um*, *ah*), as opposed to multi-word ones (e.g., *let's see*, *you know*, *I mean*). Although a majority of these instances come from the dialogue portions of the novels, fiction authors usually attempt to create natural dialogues in their prose, and thus their characters' conversational styles arguably reflect, to some extent, their own styles. However, because this measure may reflect an author's stylistic choice rather than a cognitive decline, its results should be interpreted cautiously and accepted as valid only if it significantly correlates with other lexical measures.

4.2 Syntactic Measures

The majority of our syntactic measures operate on syntactic parse trees, one for each sentence in a text. The process of generating parse trees is described in section 4.3.

S1: Syntactic complexity

Syntactic complexity is assessed by the following metrics, which have been shown to be sensitive to the effects of aging (Cheung and Kemper, 1992) and used in several studies into linguistic changes in older adults. A higher score in any of these measures indicates greater complexity.

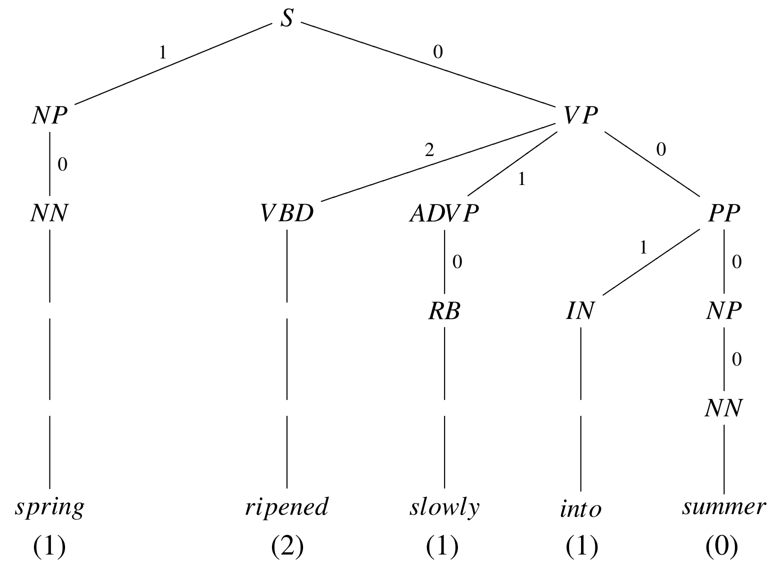
Mean Number of Clauses per Utterance (MCU) and Mean Length per Utterance (MLU):

To compute the mean number of clauses (main, subordinate, and embedded), we counted, for each parse tree corresponding to a sentence, the number of components tagged as *S* (a declarative clause) or *SQ* (a clause with subject-verb inversion), then took the average over all sentences in a text. For MLU, we simply computed the average number of words per sentence over each entire novel. Contractions, such as *isn't* and *they're*, count as two words.

Unweighted Parse Tree Depth and Yngve Depths:

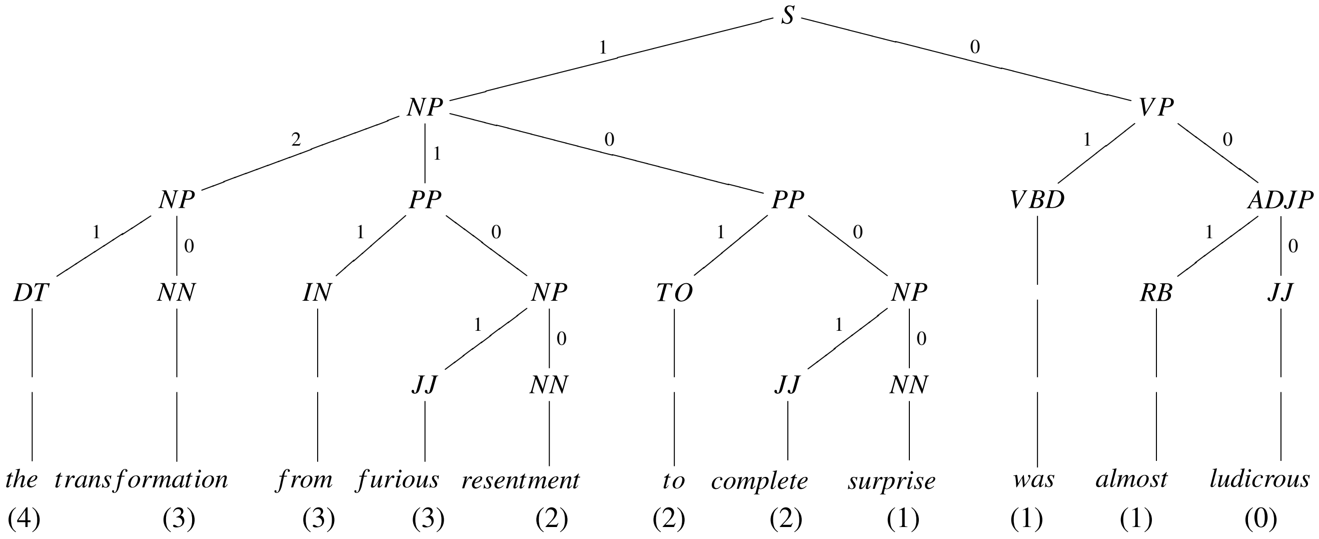
The unweighted parse tree depth measure computes the maximum unweighted depths of the parse trees corresponding to the sentences in each complete novel, and reports the average depth. This reflects the average number of embedded structures in a sentence, in order to approximate syntactic complexity, based on the assumption that deeply nested levels of embedding are associated with complex sentences. One drawback of this simple measure is the equal weight it assigns to left-branching and right-branching structures, while, given the nature of the English language, left-branching structures are more complicated and put a heavier requirement on working memory (Kemper et al., 2001). Therefore, in addition to the unweighted tree depth, we used an asymmetric measure that compensates for left-branching structures, namely, the Yngve measure, computed for both maximal and total depths (Yngve, 1960). Yngve scores are assigned incrementally, starting at 0, to the branches of each node from right to left. The Yngve depth of each word token is the sum of all the branches that connect the token to the root node. The maximal Yngve depth of each sentence is the maximum of its token depths; the total Yngve depth is the sum of all token depths. Besides tree depth, the total Yngve depth measure also takes into account the breadth of the parse tree, which corresponds to the sentence length. Thus a sentence that branches to the left will receive a higher maximal Yngve score than one that branches to the right, and may also receive a higher total Yngve score, despite having the same unweighted depth. The reported results for each text are the averages of the maximal and total Yngve depths over all sentences of the text.

The following example from P. D. James's *Cover Her Face* illustrates how the depth measures are computed on the same parse tree. Each number in parentheses indicates the Yngve depth of the corresponding word token.



Unweighted depth = 4;	Maximal Yngve depth = 2;	Total Yngve depth = 5
(at *summer*)	(at *ripened*)	$(1+2+1+1+0=5)$

In contrast to the above right-branching sentence of low complexity, the following parse tree represents a more complex, left-branching sentence from the same novel. The two sentences have the same unweighted depth, but the maximal Yngve depth for the left-branching sentence is higher. The total Yngve score of the second sentence is also higher, reflecting its left-branching structure and higher word count.



Unweighted depth = 4;   Maximal Yngve depth = 4;   Total Yngve depth = 22

D-Level:

Constructed by Rosenberg and Abbeduto (1987), the D-Level scale is a psycholinguistics-based ranking of sentences, which consists of seven levels sorted by increasing syntactic complexity for different sentence types. Reflecting the developmental stages observed in children's language acquisition, D-Level presupposes the hypothesis of linguistic regression in aging adults and was originally used to measure the linguistic competence of mildly retarded adults. The scale has been shown to correlate with measures of working memory (Kemper and Sumner, 2001), and to be sensitive to the effects of age (Cheung and Kemper, 1992). D-Level has been used extensively in recent studies into effects of AD on language production, in order to assess the extent of grammatical deficit caused by the disease.

Some limitations of the original D-Level scale are that, first, it does not account for all sentence types (Cheung and Kemper, 1992) and, second, its ordering of some types and its criterion for Level 7—"more than one kind of embedding in a single sentence"—might not correctly reflect the natural stages of language acquisition (Covington et al., 2006), leading to an incorrect model of syntactic complexity. The first limitation, in particular, renders the original scale unsuitable for our purpose, that is, assessing the average syntactic complexity of texts. To rectify this, Cheung and Kemper (1992) added Level 0 to account for simple, one-clause sentences, which are unrated in the original scale. Covington et al. (2006) proposed further modifications, which included: adding elliptical sentences and fragments to Level 0; ranking questions at the same level as the corresponding declarative sentences; adding several new structures and rearranging some existing structures based on psycholinguistic evidence; and modifying the definition of Level 7 to require two different levels of embedding. Our implementation of the syntactic complexity analyzer, described in section 4.3, is based on this revised version of the D-Level scale.

The fact that the target texts are novels, which inevitably contain dialogues, presents a complication that may affect the measures of syntactic complexity given above. Most fiction writers try to capture the essence of natural, real-life conversations in their dialogues; since spoken language tends to have lower complexity, with shorter sentences, fewer embedded clauses, less complex grammar, and more fragments, the proportion of dialogue in each novel partly determines its complexity scores. An optimal solution is to perform separate syntactic analysis on the dialogue portions and the narrative portions; however, this separation of dialogue from narrative cannot be accomplished, given the properties of our scanned texts (see Appendix B.1 for a detailed discussion of the problems). Consequently, the results of the S1 measures might not reflect the absolute syntactic levels of the authors.

S2: Passive voice

We approximated the frequency of passive voice usage by counting the number of sentences containing a *be*-passive, a *get*-passive or a past participle verb followed by a *by*-phrase. Bare passives (those not headed by *be* or *get*—such as the verb *headed* in this clause), often cannot be distinguished from the perfect use of past participles if not accompanied by a *by*-phrase. Consider the following examples (Huddleston and Pullum, 2002):

i Considered by many overqualified for the post, she withdrew her application.

ii Now fallen on hard times, he looked a good deal older.

The underlined past participle in [i] is a passive verb (as in "She was considered by many to be overqualified for the job"), whereas the one in [ii] is used in perfect tense ("He had fallen on hard times") without being explicitly marked by the auxiliary *have*. If the phrase *by many* is omitted, the two examples become identical in form and cannot be distinguished without considering the semantic content. Another complication, perhaps even more severe, is distinguishing cases in which past participles are used as adjectives from bare passives. The following examples illustrate the problem:

iii He was fired.
iv He was drunk.
v He was pleased.

In these examples, *fired* is a passive verb, *drunk* takes on adjectival function, while *pleased* is ambiguous: depending on the semantic context, it can be a passive verb ("He was pleased by her compliments.") or an adjective ("He was pleased with himself."). Our implementation, relying solely on syntax, detects only the explicit forms of passive (in which the verb phrase is headed by *be* or *get*) and some forms of bare passive (those followed by a *by*-phrase). The measure reports the percentage of sentences containing these passive forms over the total number of sentences, as well as the percentages of *be*-passives, *get*-passives, or passives with a *by*-phrase over all passive sentences. (It is worth noting that, since a passive sentence may contain both a *be*-passive and a *get*-passive, the percentages of *be*- and *get*-passives for each novel do not necessarily sum to 100%.)

4.3 Implementation Details

Given a plain text file, the process of analyzing lexical and syntactic features consists of the stages summarized in Table 3, which will be addressed individually in this section.

Table 3: Program overview

1.	Separate punctuation marks and clitics from word tokens.
2.	Determine sentence boundaries.
3.	Generate a parse tree for each sentence.
	Fix incorrect tags made by the parser.
4a.	Disambiguate word sense and determine WordNet depth.
4b.	Match each pattern against a parse tree.

1. Separating punctuation marks and clitics:

We simply added spaces before and after every punctuation mark and clitic, which can be a contraction (e.g., *I’m*, *they’ve*, *isn’t*) or a genitive marker (e.g., *John’s*), to separate them from the word tokens to which they are attached. For instance,

> ' Once upon a time there were three little girls — '
>
> ' Oh look what he ’s doing now ! '
>
> ' And their names were — '
>
> ' Come here , come here . '
>
> ' And they lived at the bottom of a well . ' [M19]

The output of this stage is used as input data for the vocabulary (L1) and lexical repetition (L2) measures.

2. Determining sentence boundaries:

We used a rule-based, deterministic algorithm to identify boundaries among sentences. When periods, question marks and exclamation marks are encountered, they are assumed to mark the end of a sentence. The algorithm then considers predefined exceptions, such as: in the case of a period, whether it is a part of an abbreviation that rarely or never ends a sentence (e.g., *Mr.*, *Mrs.*, initials followed by a last name, such as *P. D. James*, or *St.* as an abbreviation for *Saint*, but not when it stands for *Street*); in the case of a question mark or an exclamation mark, whether it is followed by a quotation mark and a lower case word or a proper name, which may indicate direct speech or inner thought (in the latter case, the quotation mark is often omitted). Finally, every complete sentence is marked with XML-style markers, <s> … </s>.

In the following examples, the exclamation mark and the question mark are disqualified as sentence-ending markers.

> <s> ' It ’s like philosophy ! ' Harvey had exclaimed at one point . </s> [M19]
>
> <s> " We do not agree , eh ? " said Poirot . </s> [C1]

In addition, we also consider ellipses and em dashes as potential end punctuation marks, since they are sometimes used to indicate interruptions, hesitation, or thoughts trailing off, both in dialogue,

<s> ' Well , no , I mean I do n't think you — ' </s>

<s> ' Never mind , we can talk later. </s> [M19]

and in narrative,

<s> Supposing … </s>

<s> She choked her fears down bravely . </s> [C2]

At this stage, the data is ready for the mean sentence length measure (S1).

3. Generating parse trees:

We used the Charniak reranking parser (Charniak, 2006) to process the sentences produced in the previous step. Out of the several parsers we tested, the Charniak parser yielded the highest accuracy for the type of data used in our analysis; however, it does, occasionally, make errors in determining part-of-speech or levels of embedding. We have identified some patterns of error the parser often makes, the most critical of which is that all conjugated forms of the verbs *be*, *have*, and *do* are tagged as auxiliaries (AUX or AUXG), even when they are in fact content verbs. To correct this, we wrote a script that proceeds down each parse tree and marks the last AUX- or AUXG-tagged instance of *be/have/do* in each verb-phrase branch followed by a complement (a noun phrase, an adjective phrase, or a prepositional phrase) as a content verb of the appropriate tense. For example, `(VP (AUX is) (ADJP (JJ busy)))` becomes `(VP (VBZ is) (ADJP (JJ busy)))`; `(VP (MD will) (VP (AUX be) (ADJP (JJ free))))` becomes `(VP (MD will) (VP (VB be) (ADJP (JJ free))))`; `(VP (AUX are) (VP (VBN set) (ADJP (JJ free))))` remains unchanged, since *are* in this case is not a content verb.

Ellipses involving *be* present a challenge; consider the following:

I asked Joan if she was coming, and she said she was$_1$.

I asked Joan if she was a musician, and she said she was$_2$.

The subscripted instances of *be* in the above examples are fixed as follows: the first sentence contains a verb phrase ellipsis, thus *was*$_1$ is correctly tagged as an auxiliary, while the second sentence contains a noun phrase ellipsis, hence *was*$_2$ is a content verb. However, for more complicated sentences with several branches of verb phrase, resolving the ellipsis binding might not be computationally straightforward. Furthermore, whether an elliptical construction involves *be* as a content verb or an auxiliary has little or no impact on the results of our parse-tree-based measures. Because fixing this type of error yields only marginal gain, we left the incorrectly tagged instances of *be* in elliptical constructions unchanged.

The output of this stage becomes the data for measures that operate on syntactic parse trees or part-of-speech tags of word tokens, including the verb specificity measure (L3), word class proportion (L4), fillers (L5), mean number of clauses per sentence (S1), and average parse tree depths (S1).

4a. Disambiguating word sense and measuring WordNet depth:

This procedure applies only to the noun specificity measure, which relies on WordNet depths as an approximation of specificity ranks. To determine the WordNet synset of a noun token in context (that is, the sense being used), we ran WordNet::SenseRelate::AllWords, a word-sense disambiguation program by Pedersen (2009), over individual sentences from each text. Since there may be more than one path from the root to a synset (for instance, the first synset of *dog*, denoted dog_1, as shown below), the measure computes both the minimum depth and the maximum depth for each identified synset, and reports the two corresponding types of average separately.

$\text{entity}_1 > \text{physical\_entity}_1 > \text{object}_1 > \text{whole}_2 > \text{living\_thing}_1 > \text{organism}_1 > \text{animal}_1$
$> \text{domestic\_animal}_1 > \mathbf{dog_1}$

$\text{entity}_1 > \text{physical\_entity}_1 > \text{object}_1 > \text{whole}_2 > \text{living\_thing}_1 > \text{organism}_1 > \text{animal}_1$
$> \text{chordate}_1 > \text{vertebrate}_1 > \text{mammal}_1 > \text{placental}_1 > \text{carnivore}_1 > \text{canine}_2 > \mathbf{dog_1}$

4b. Matching patterns and parse trees

The remaining syntactic measures—D-Level score (S1) and passive proportion (S2)—use a pattern-matching algorithm, which assigns scores to sentences that match predefined patterns. Each pattern describes the structural, syntactic and, optionally, lexical properties required in a matching parse tree. The set of patterns for each measure is included in Appendix B.3; in this section, we describe the mechanism of this pattern-matching algorithm.

Overview:

For each sentence type specified in the revised D-Level scale and each passive structure, we defined one or more pattern that describes the necessary and sufficient conditions for a matching parse tree (each parse tree corresponds to one sentence in the datasets). When a match is found, the algorithm assigns a score to the sentence. In the case of D-Level, this score is a value between 0 and 7, corresponding to the eight levels of the scale. For passive proportion, the score is binary, 1 for a tree containing an identified passive structure, and 0 otherwise.

Two modes of pattern-matching were implemented: root-match (the pattern has to match the tree from the root node) or branch-match (the pattern can match any subtree). We also implemented a set of special pattern symbols to specify the exact match location in a parse tree.

Language:

Our syntactic complexity analyzer was originally written in Scheme, which offers the following advantages: first, recursing down a parse tree is simple, given the nature of the language, and second, the parse tree format (determined by the parser) is conveniently a well-defined nested list in Scheme. While our Scheme program works well, speed and scalability are its major drawbacks. The program was thus rewritten in object-oriented Python. A preprocessing stage becomes necessary, in which the input parse trees and patterns, read as "flat" strings, are transformed into levelled data structures. The Python program, retaining all of the functionalities of the original Scheme counterpart, offers the ability to process multiple input files in batch mode, quickly and conveniently—an important improvement, considering the size of our datasets.

Parse Tree Format:

A parse tree can be defined recursively as either a leaf node, which consists of a tag and a value, or a non-leaf node, which consists of a tag and a list of other nodes (i.e., its subtrees). The format we use reflects this recursive structure, and is also the standard output format of most parsers. The tags are standard part-of-speech tags used by the Penn Treebank (with some additional tags used by the Charniak parser), and the values are simply word tokens.

Leaf node:	(*tag value*)
Non-leaf node:	(*tag* $child\text{-}node_1$ $child\text{-}node_2$ … $child\text{-}node_n$)

The following is an example of a well-formed parse tree (*):

```
(S (NP (PRP This)) (VP (VBZ is) (NP (DT a) (JJ simple) (NN example))) (.  .))
```

Basic Pattern Format:

A basic pattern has the same format as a parse tree, with the exception that the values of its leaf nodes can be omitted. More specifically:

Leaf pattern 1:	(*tag value*)
Leaf pattern 2:	(*tag*)
Non-leaf pattern:	(*tag* $child\text{-}pattern_1$ $child\text{-}pattern_2$ … $child\text{-}pattern_n$)

The parse tree (*) given above is a pattern itself, and so are the following, with varying degrees of specificity:

```
(S (NP (PRP)) (VP (VBZ) (NP (DT) (JJ) (NN))))

(S (NP this) (VP (VBZ is) (NP (NN))))

(VP (VBZ) (NP (NN)))

(NN)
```

A set of special symbols is available to specify further requirements on a matching parse tree; the parse tree format containing these symbols is presented later in this section.

Basic Pattern Matching Rules:

At the leaf level, a leaf pattern matches a leaf node if they have the same tags and the same values. Value matching is case-insensitive, while tags must be an exact match. If the value of the leaf pattern is omitted, then only the tags are considered—the pattern accepts any value at the corresponding location in the parse tree. At the non-leaf level, informally, a pattern matches a parse tree from the root if they have the same tags, and each child node of the pattern matches, in order, the corresponding node in a subset of the child nodes of the parse tree, which are not necessarily adjacent (sibling) nodes. We define this last concept formally as follows.

> Assume that T_i's are well-formed parse trees, and P_j's well-formed patterns for all $i \in [1,n]$ and $j \in [1,m]$ such that $m \leq n$. Let t_{T_0} and t_{P_0} be two node tags; then $T_0 = (t_{T_0}\ \ T_1\ \ T_2\ \dots\ T_n)$ is a well-formed parse tree, and $P_0 = (t_{P_0}\ \ P_1\ \ P_2\ \dots\ P_m)$ a well-formed pattern. P_0 matches T_0 from the root if:
>
> - t_{T_0} and t_{P_0} are the same, and
> - there exists a set $T_{a_1},\ T_{a_2},\ \dots\ T_{a_m}$ such that $a_i \in [1,n]$, $a_1 < a_2 < \ldots < a_m$, and P_i matches T_{a_i} for each i $\in [1,m]$.

In the second mode of matching (branch-matching), a pattern can match a parse tree either from the root node or from an embedded node at any sub-level down the parse tree according to the same root-matching rules.

The basic pattern examples, given previously, all match the parse tree (*) in branch-matching mode, while only the first two patterns match from the root. On the other hand, the parse tree does not match any of the following patterns in either mode, because of some mismatched components, which can be tags, values, or levels of embedding. These components are underlined.

```
(S (NP (NNP)) (VP))

(S (NP) (VP (VBZ) (NP (DT the) (JJ) (NN))))

(S (VBZ))

(RB)
```

Special Symbols:

The basic pattern-matching rules let us specify the exact tags, values, and embedding levels required in a parse tree; however, apart from the optional value specification, matching operates on a literal basis. We introduced two types of special symbols, which add flexibility to the pattern-matching algorithm, analogous to the power of regular expression over literal string matching:

- **Content symbols** can replace node tags or leaf values. If a single underscore (“_”) replaces a node tag, this matches any tag at the corresponding location in the parse tree. If a square-bracketed list of node tags (or leaf values) is encountered, any of these tags (or values) can match the tag (or value) at the corresponding location in the parse tree.

- **Structural symbols** are optionally added in front of a child-node pattern to specify additional information about the syntactic structure of the match. The implemented syntactic symbols (described in Table 4) are either unary or binary. The format of a non-leaf pattern now becomes (with square brackets indicating optional arguments):

$$(\textit{tag}\ \ [\textit{unary-symbol}_1]\ \ \textit{pattern}_1\ \ [\textit{binary-symbol}_{2,3}]\ \ \textit{pattern}_2\ \ \textit{pattern}_3\ \ldots)$$

Examples using these special symbols, both content and structural, are given in Table 5. In these examples, root-matching mode is assumed.

Table 4: Structural symbols for pattern matching

Symbol	*Type*	*Effect*
+	unary	the match must occur at the current location in the parse tree
~	binary	the first pattern is optional and may or may not occur at the current level, but the second pattern is required
-	unary	the pattern must not match the remaining subtrees at the current level
-	binary	the first pattern must not occur at the current level until the second pattern is encountered
*	unary	the match can occur at any sub-level
^	unary	the match can occur at any sub-level as long as the path to that sub-level contains only the specified tag

Table 5: Examples of patterns containing special content and structural symbols

	Pattern	Imposed Requirement(s) on Parse Tree	Example Matches
1.	(NP (_ sheep))	Matches any noun phrase (NP) that has a child node whose value is “sheep,” without any requirement on its tag. Thus “sheep” can be a singular noun (NN), or a plural noun (NNS), or any other tag.	(NP (NN sheep)) (NP (JJ black) (NNS sheep))
2.	(NP ([NN NNS] sheep))	Compared to #1, this pattern restricts the choice of tags to exactly two options: NN or NNS. Any other tag will be rejected.	(NP (DT a) (NN sheep)) (NP (NNS sheep))
3.	(NP (NN [sheep wolf]))	Allows choices in leaf values. Only noun phrases with singular noun “sheep” or “wolf” are accepted.	(NP (NN sheep)) (NP (DT that) (NN wolf))
4.	(NP +(NN))	Requires that the first child node of NP be NN.	(NP (NN sheep)) ~~(NP (DT the) (NN sheep))~~
5.	(NP (DT) +(NN))	Requires that NP have a DT child node immediately followed by an NN node. DT and NN must be immediate siblings; if another node exists in between, the parse tree is rejected.	(NP (DT the) (NN sheep)) ~~(NP (DT the)~~ ~~((JJ black) (NN sheep))~~
6.	(NP (DT) ~(JJ) (NN))	Allows the JJ component to be optional. This pattern matches a noun phrase that contains a DT node followed by a NN node, with zero or more JJ nodes in-between. If any component other than JJ exists between DT and NN, the parse tree is rejected.	(NP (DT the) (JJ sad) (JJ black) (NN sheep)) ~~(NP (DT the) (CD three)~~ ~~(JJ black) (NN sheep))~~
7.	(NP -(JJ))	Requires that NP not contain any JJ child node.	(NP (CD two) (NNS wolves))
8.	(NP (DT) -(JJ))	Requires that NP not contain a JJ child node after DT.	(NP (DT the) (NNS wolves))
9.	(NP (DT) -(JJ) (NN))	Requires that no JJ components exist among the child nodes of NP after DT and before NN.	(NP (DT the) (NN Attorney) (JJ General))
10.	(NP *(JJ))	Allows JJ to be at any sub-level relative to NP. In the matching tree example, the path from the root to JJ is: NP > ADJP > JJ.	(NP (DT a) (ADJP (RB very) (JJ sad)) (NN sheep))
11.	(NP ^(JJ))	Similar to #10; however, this requires JJ to be embedded only within NP, thus the example match in #10 is rejected. For this tree to be accepted, we can specify tag options, which allows JJ to be embedded within either NP or ADJP: ([NP ADJP] ^(JJ)).	(NP (NP (JJ sad) (NNS sheep)) (CC and) (NP (JJ hungry) (NNS wolves)))

5 Results

In this section, we present and discuss the results of our experiments conducted on novels by Iris Murdoch, Agatha Christie, and P. D. James. For all lexical measures, the outlier in Christie dataset, *Passenger to Frankfurt*, is excluded from the overall trend, but is included in the graphs as a single datapoint to demonstrate the effects of Christie's research on her linguistic properties. Among Murdoch's novels, *The Italian Girl* has an unusually low word count (48,448) relative to the average of the remaining 19 Murdoch novels (138,312) and is therefore excluded from the measures that are normalized in length, to keep the length threshold at 55,000 word tokens for all three authors. None of the novels is considered an outlier when syntactic markers are measured.

Simple linear regression was performed on each set of results and is included in the graphs to illustrate the overall increasing or decreasing trend. Each regression model was tested for statistical significance; a probability of 0.05 or below, equivalent to a confidence level of at least 95%, is required to reject the null hypothesis that the model is unfit to represent the datapoints. Correlation between measures were tested using the Spearman rank-order correlation coefficient method.

5.1 Lexical Results

Lexical marker L1: Vocabulary Size

Change in vocabulary size was assessed by two measures: type/token ratio (TTR) and word-type introduction rate. Figure 1 displays variations in the TTRs of each author over time. The TTRs of Murdoch's novels fluctuate slightly before the 50-year mark, begin to rise in her 50s, peaking in her mid-60s at *The Sea, the Sea*, before plummeting to a trough with her last novel. Although the entire dataset exhibits a statistically insignificant increasing trend [$F(1,19) = 0.19$, $P = 0.6651$], Murdoch's first 15 novels, excluding *The Italian Girl*, show a significant rise [$F(1,13) = 13.41$, $P = 0.0029$], while in the last 5 novels, the decreasing trend is steeper and also significant [$F(1,3) = 14.17$, $P = 0.0328$]. The TTRs of Christie's novels fluctuate in the 0.07 to 0.084 range before her early 60s, and begin to drop from that point on, reaching a bottom at her second last novel, *Elephants Can Remember*. A significant decline [$F(1,13) = 9.29$, $P = 0.0093$] is found for the entire period, excluding *Passenger to Frankfurt*, which is clearly an outlier in this dataset, breaking away from the overall decline and higher than any of Christie's other novels. On the other hand, James's TTRs vary in the 0.09 to 0.11 range without apparent signs of decline; however, the irregularity of the values makes the slight rising trend insignificant [$F(1,13) = 0.59$, $P = 0.4550$].

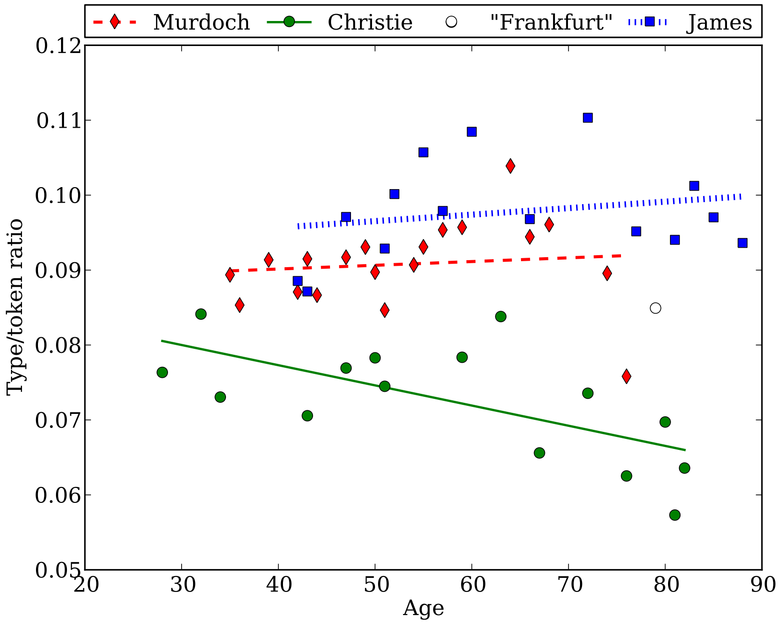
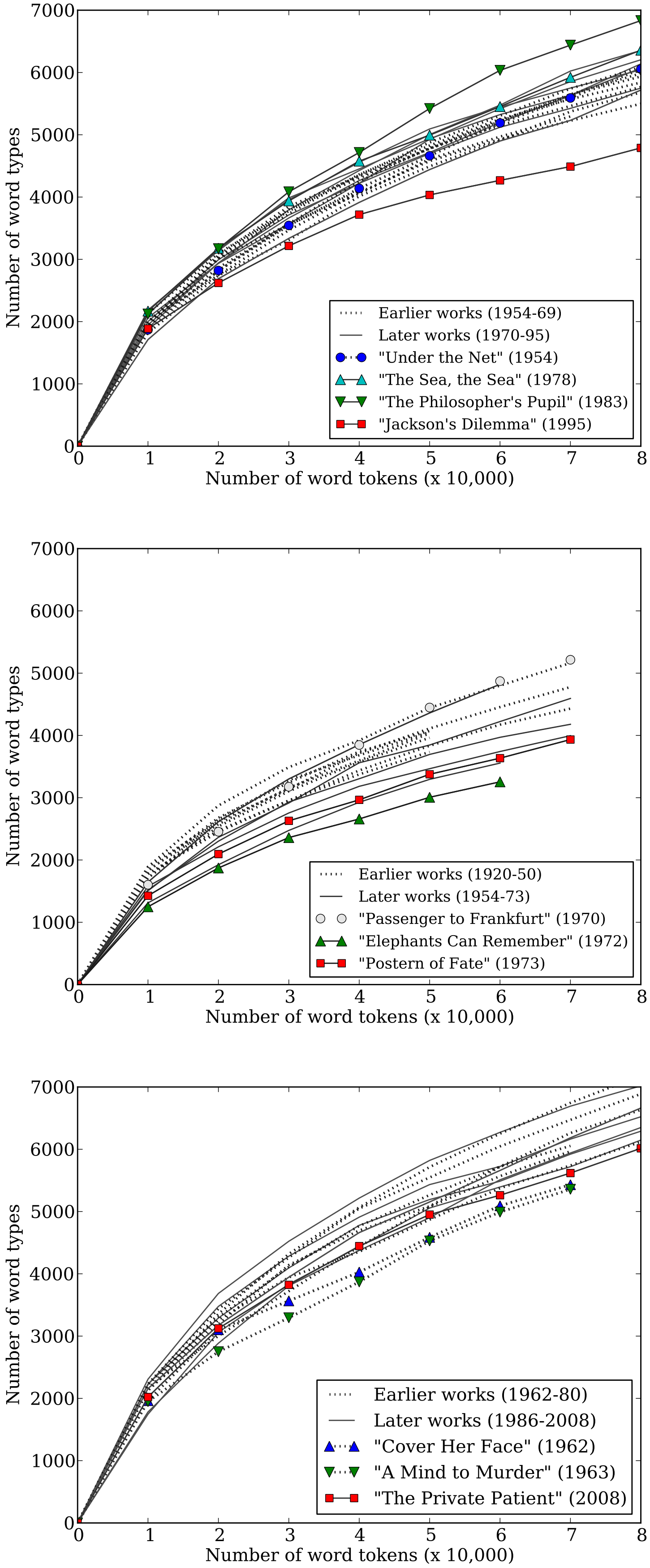


Figure 1: Type/token ratio within the first 55,000 tokens

These results are reinforced by the word-type introduction rates (WTIR), shown in Figure 2. In these graphs, each line corresponds to a novel, reflecting the vocabulary growth (i.e., the number of unique word types) measured at every 10,000 word-token interval. The lines may overlap, and a clustering of lines indicates higher consistency than a more scattered set of lines. The novels by each author are divided into two groups, the first represented as dotted lines and the second as solid lines. For Christie, this division coincides with her change of writing method, from the typewriter to the dictaphone. Her novels, containing from 50,000 to under 80,000 word tokens, are measured up to a maximum of 70,000 tokens. To ensure a fair comparison, Figures 2a and 2c are scaled to focus on the first 70,000-token portions; the complete graphs are shown in Figure 3 (p. 50).

Murdoch's last novel, *Jackson's Dilemma*, stands out with an unusually low rate of vocabulary growth compared to her previous works, all but one of which cluster together in a concentrated band. This confirms the TTR results that the decline in Murdoch's vocabulary occurred abruptly,



(a) Iris Murdoch

Rates of vocabulary growth are consistent, except for *Pupil* and *Jackson*. The latter, which is also Murdoch's last novel, has the lowest rate.

(b) Agatha Christie

Gradual decline. Earlier novels have higher rates; most later novels have lower rates, with the second last novel, *Elephants*, being the lowest.

(c) P. D. James

The rates are mostly consistent, with earlier and later works intertwined. The first two novels have the lowest rates of vocabulary growth.

Figure 2: Word-type introduction rate up to the 70,000th token

and is consistent with Garrard et al. (2005) in that this decline is evident in *Jackson's Dilemma*. Employing similar methods, our longitudinal approach reveals additionally that the decline became severe *while* she was writing this last novel: Figure 3a shows that the vocabulary growth of *Jackson's Dilemma* begins to slow down significantly only *after* the 40,000th token, compared to the majority of Murdoch's works. A similar declining tendency, though more gradual, can be seen in Christie's last two novels, *Elephants Can Remember* (which has the slowest rate of growth) and *Postern of Fate*. All of Christie's earlier works stay in the upper range, while most of the later works (except for *Destination Unknown* and *The Clocks*) occupy the lower range, indicating a progressive impoverishment of vocabulary. For P. D. James, a different picture emerges: the rates stay relatively consistent, with earlier works and latter works intertwined, apart from her first two novels, which stand out in a slightly lower range. Her last novel remains in the mid range up to the 50,000-token mark, then converge towards her lower range from the 60,000-token mark onwards, but does not greatly depart from her usual rates.

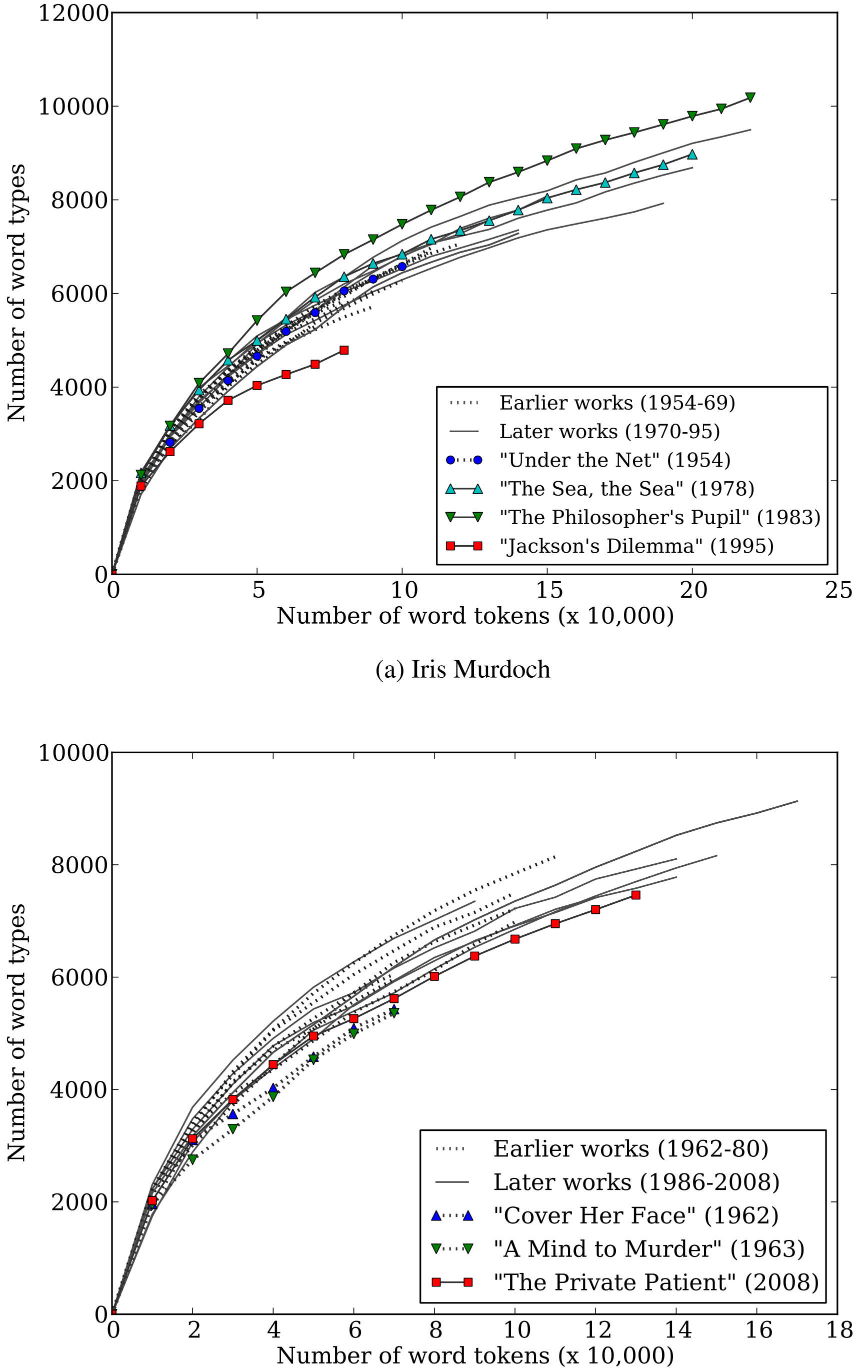
Statistical tests comparing increasing-sized portions among all the novels confirm a sharp decline in the word-type introduction rate in Christie's works over time [$P < 0.0083$ for blocks of up to 50,000 words]. No significant trends are found for Murdoch and James, which is not surprising: unlike Christie's graph, which can be divided into two portions for the earlier and later novels (with two exceptions), there are no obvious longitudinal patterns for the other authors.

Table 6: Correlation between vocabulary measures

WTIR up to token:	10,000	20,000	30,000	40,000	50,000
TTR of Murdoch	+0.66	+0.74	+0.84	+0.92	+0.98
TTR of Christie	+0.78	+0.95	+0.95	+1.00	+1.00
TTR of James	+0.75	+0.80	+0.89	+0.94	+0.97

(All correlations have P-value < 0.01)

Table 6 shows the correlation between TTR and WTIR measured at various points in each text. A very strong correlation with high significance is found when WTIR is evaluated at the 50,000th token, even when many of Murdoch's and James's novels fall between 100,000 and 220,000 in token count. The results of both measures highlight the fact that Murdoch's last novel and Christie's last two share a common characteristic: their vocabulary sizes deviate from the norms set by the authors' earlier works. Murdoch's decline is abrupt, while Christie's is more gradual over time.



(a) Iris Murdoch

(b) P. D. James

Figure 3: Word-type introduction rate (complete texts)

Lexical marker L2: Lexical repetition

Figure 4 shows the proportion of lexical repetitions within 10 subsequent content words, computed over the number of all content words in each novel.

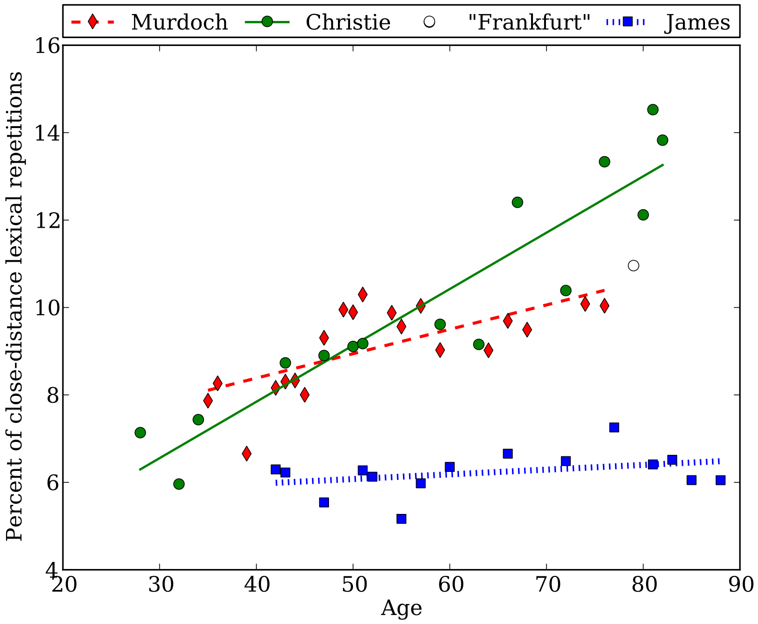


Figure 4: Lexical repetitions within 10 subsequent content words

Murdoch's overall trend is a significant increase, which peaks at the 51-year mark. Christie's repetition rates show an even steeper rise with high certainty, with the highest rates in her last two novels, of which 14.53% and 13.83% of the content word tokens are repeated within close distance, in sharp contrast to the rates of 7.14% and 5.96% in her first two novels. On the other hand, the repetition rates in James's novels remain relatively stable in the low range (5.54 to 7.26%).

Table 7: Statistical significance test results for lexical repetition measure

	MURDOCH		CHRISTIE		JAMES	
	Coeff.	$F(1, 18)$	Coeff.	$F(1, 13)$	Coeff.	$F(1, 13)$
Distance 10	0.0558	15.99^{**}	0.1289	83.46^{**}	0.0108	1.94
Distance 20	0.0526	7.90^{*}	0.1535	63.58^{**}	0.0153	1.90

$^{*}P < 0.05$ $^{**}P < 0.01$

When the distance is extended to 20 subsequent content words, similar patterns of changes are observed in all three authors, as indicated by the coefficient values in Table 7. The rising trend is more pronounced in Christie's results, whereas Murdoch's trend becomes slightly less steep, but both with significance. James's marginal increase is again insignificant.

Table 8: Correlation between lexical repetition and vocabulary measures

LR20	TTR	WTIR 50,000
Murdoch	-0.14	-0.11
Christie	-0.79^{**}	-0.79^{**}
James	-0.52^{*}	-0.45

$^{*}P < 0.05$ $^{**}P < 0.01$

Table 8 displays the correlation between rates of lexical repetition of distance 20 (LR20) and the two vocabulary measures, TTR and WTIR at 50,000 tokens. As predicted, repetition rate is negatively correlated with vocabulary size, although only Christie's results show a strong and highly significant correlation. A milder negative correlation with significance is found between James's lexical repetition and TTR. The rise in Murdoch's repetition rates after the 60-year mark coincides with the drop in TTR, while the earlier portions are not as well-correlated. It is notable, however, that at the 51-year mark (*A Fairly Honorable Defeat*), Murdoch's repetition rates climb to a peak at both distances 10 and 20, while her TTRs reach a low at the same time.

Lexical marker L3: Word specificity

Word specificity is assessed by two measures: the percentage of high-frequency, low-specificity verbs, and the average WordNet depth of nouns.

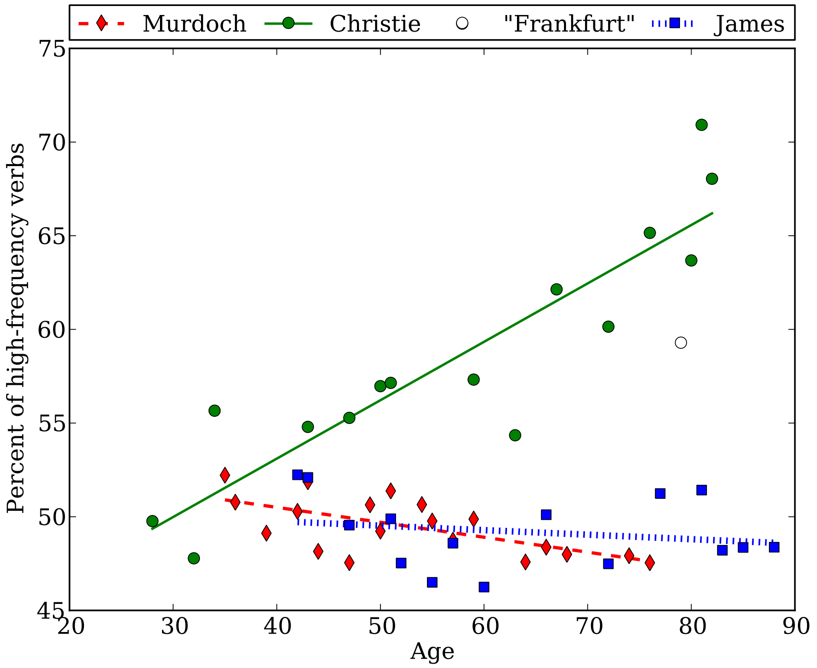


Figure 5: Proportion of high-frequency verbs within the first 55,000 tokens

Figure 5 displays the percentages of verbs that belong to the verb list given in section 4.1, which contains thirty-five high-frequency verb types, measured within the first 55,000 tokens of each novel. Christie's rates reveal a marked increase, from a low of 48% at her 1922 title, *The Secret Adversary*, to a high of 71% at her second last novel, *Elephants Can Remember* [$F(1,13) = 55.74$, $P < 0.0001$]. That thirty-five verbs account for 71% of all verbs in a novel of nearly 62,000 words suggests a severe deficit in verbs, due to either word retrieval problems or an impoverished vocabulary. Christie's extensive research for *Passenger to Frankfurt* greatly reduced the percentage of these verbs: at 59%, *Frankfurt* has the lowest rate among Christie novels in a period of 15 years. In contrast, the results for Murdoch and James remain relatively stable below 53%. For Murdoch, a moderate decrease of high significance was found [$F(1,17) = 12.13$, $P < 0.0028$], while James's slight decreasing trend was statistically insignificant [$F(1,13) = 0.53$, $P = 0.4789$].

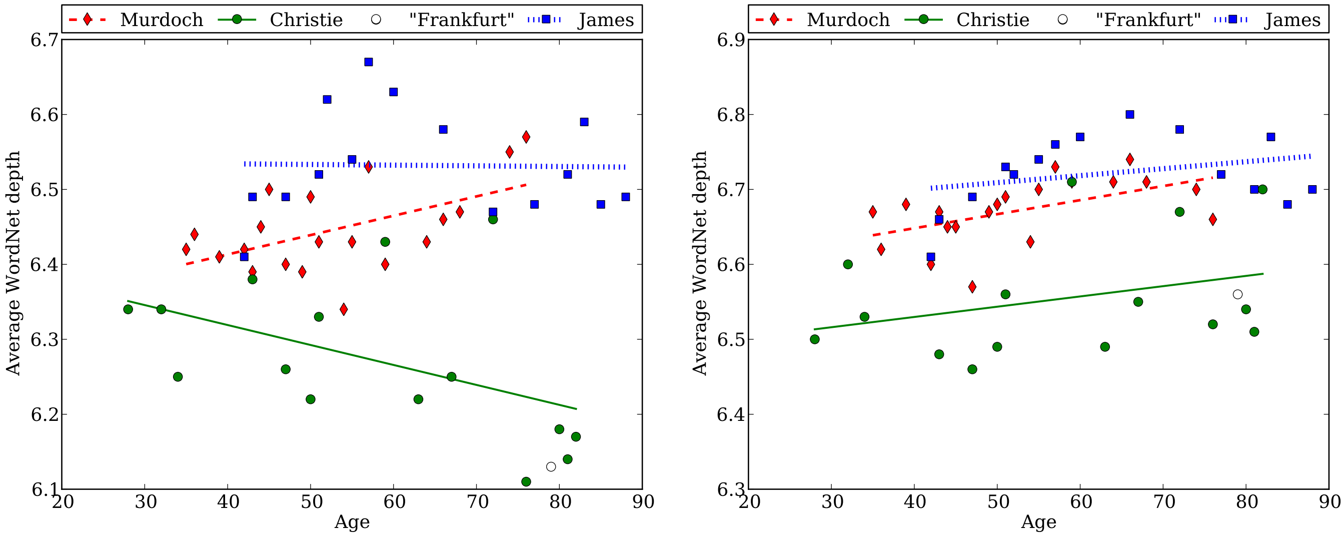
Table 9: Correlation between high-frequency verbs and other lexical measures

	LR	TTR	WTIR 50,000
Murdoch	-0.16	-0.33	-0.35
Christie	$+0.95^{**}$	-0.81^{**}	-0.81^{**}
James	$+0.60^{*}$	-0.86^{**}	-0.88^{**}

$^{*}P < 0.05$ $\quad$ $^{**}P < 0.01$

Table 9 shows this measure to be significantly correlated with the lexical repetition measure and negatively correlated with the vocabulary measures for both Christie and James. The implications of these results are that a larger vocabulary entails fewer lexical repetitions and less reliance on common verbs of low specificity.

Figures 6–7 show the results of our noun specificity approximation. As described in section 4.1, this measure computes the average WordNet depth of all noun synsets, both in terms of WordNet's maximum depth and minimum depth.



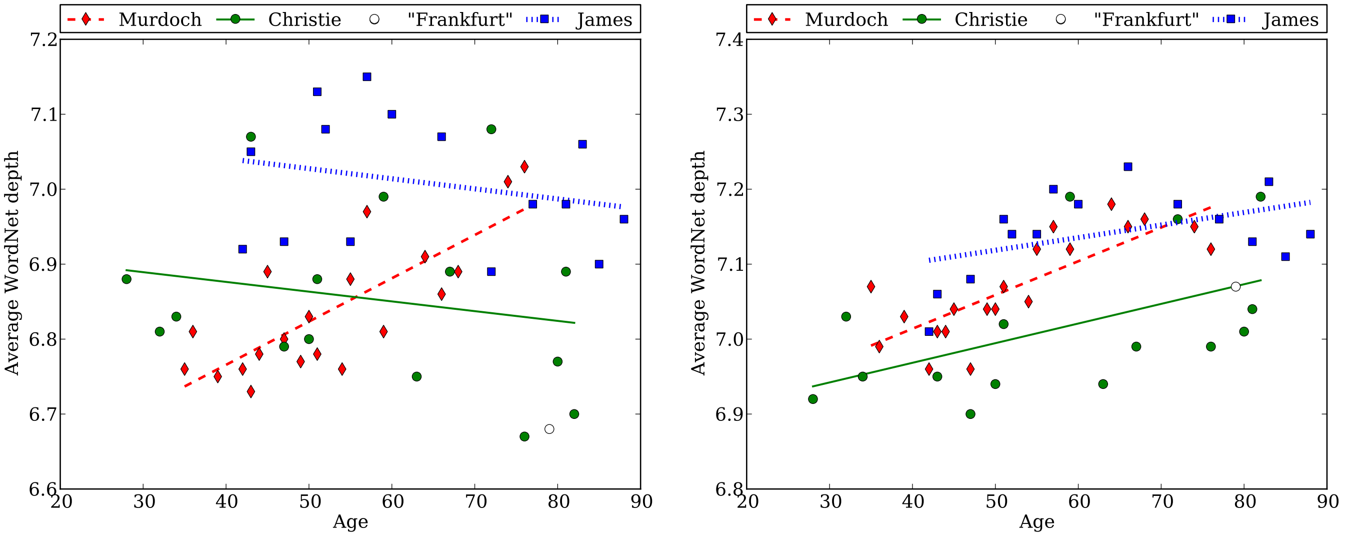
(a) by token

(b) by type

Figure 6: Noun specificity (minimum WordNet depth)

As shown in Figures 6a and 7a, Murdoch's noun specificity ranks oscillate in the early novels and exhibit an overall increasing trend for both maximum and minimum depths. Her noun type specificity ranks also fluctuate up to the late 50s, then remain relatively stable with a slight drop at her last novel (see Figures 6b and 7b). Christie's results, on the other hand, vary in a wide range, then plunge to a trough with *Endless Night* in her mid-70s, contributing to an insignificant decreasing trend overall. Surprisingly, despite the drop in noun token specificity ranks, Christie's noun type specificity remains high in the last few novels and rises sharply with *Postern of Fate*. With its irregularity, the overall trend is an increasing one with low significance. Similarly, James's noun token specificity fluctuates with no clear pattern, while her type ranks constitute a consistent rising trend up to her late 60s, followed by a gradual decline with one exception at 83 with *The Murder Room*. Statistical test results are reported in Table 10; the only significant trends are those of Murdoch's novels.

Table 11 reports the correlation coefficients between the noun specificity measures. Moderate to high correlation was found between maximum and minimum depth results. Correlation between token and type results is relatively low, as reflected in the corresponding graphs.



(a) by token

(b) by type

Figure 7: Noun specificity (maximum WordNet depth)

The wide oscillation range of Murdoch's and Christie's noun token results, the inconsistencies between type and token specificity averages and the overall irregularity lead us to question the suitability of WordNet depth as a measure of specificity. Appendix B.2 examines this issue in further detail.

Table 10: Statistical significance test results for specificity measures

	MURDOCH		CHRISTIE		JAMES	
	Coeff.	$F(1, 18)$	Coeff.	$F(1, 13)$	Coeff.	$F(1, 13)$
Low-specificity verb proportion	–0.0798	12.13^{**}	0.3117	55.74^{**}	–0.0240	0.53
Noun token (min)	0.0026	7.32^{*}	–0.0027	3.72	–0.0001	0.01
Noun token (max)	0.0058	29.73^{**}	–0.0013	0.54	–0.0013	0.85
Noun type (min)	0.0019	6.83^{*}	0.0014	1.44	0.0009	1.23
Noun type (max)	0.0045	30.68^{**}	0.0026	4.62	0.0017	3.43

$^{*}P < 0.05$ $^{**}P < 0.01$

Table 11: Correlation between noun specificity measures

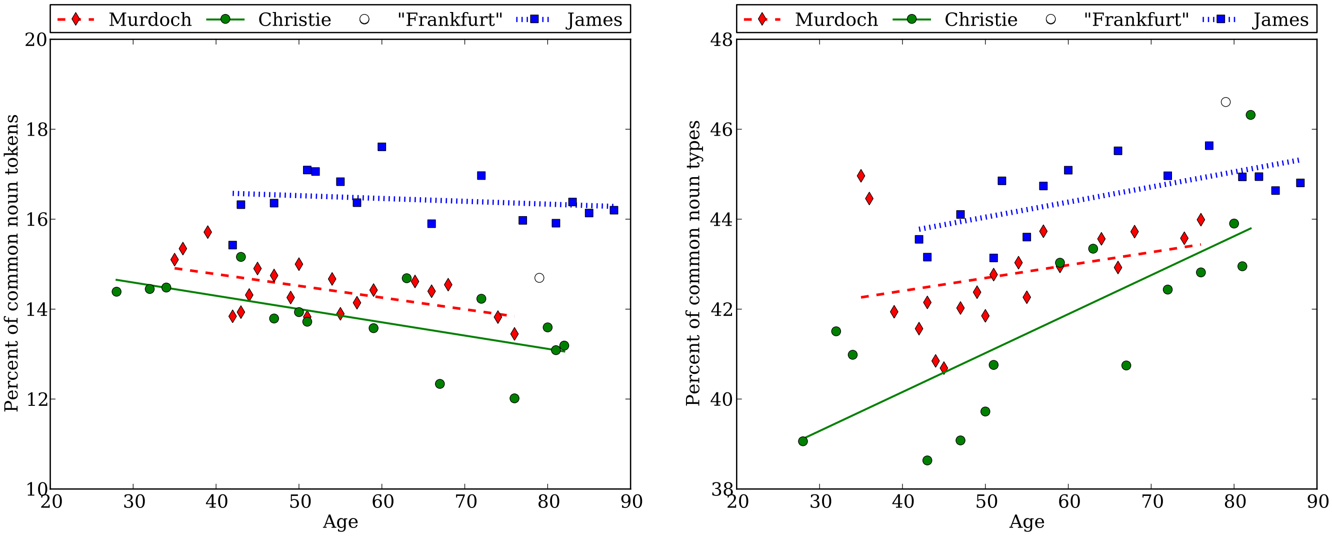
		1	2	3	4
MURDOCH	1. Noun token (max)	—	$+0.82^{**}$	$+0.68^{**}$	$+0.45^{*}$
	2. Noun token (min)		—	+0.44	+0.31
	3. Noun type (max)			—	$+0.84^{**}$
	4. Noun type (min)				—
CHRISTIE	1. Noun token (max)	—	$+0.72^{**}$	+0.25	+0.18
	2. Noun token (min)		—	+0.06	+0.20
	3. Noun type (max)			—	$+0.87^{**}$
	4. Noun type (min)				—
JAMES	1. Noun token (max)	—	$+0.82^{**}$	$+0.52^{*}$	+0.39
	2. Noun token (min)		—	$+0.57^{*}$	$+0.55^{*}$
	3. Noun type (max)			—	$+0.94^{**}$
	4. Noun type (min)				—

$^{*}P < 0.05$ $^{**}P < 0.01$

Lexical marker L4: Word class deficit

Figures 8–12 display the changes in proportions of nouns, pronouns, content verbs, adjectives and adverbs, in terms of token count and type count. Table 12 shows the results of statistical significance tests for each word class of interest, and Tables 13–14 report the correlation coefficients between the different word classes.

In contrast to Garrard et al. (2005), whose approximation of grammatical class proportion found no significant differences among three of Murdoch's novels, our analysis, using part-of-speech tagging, discovered longitudinal variations in the datapoints. Among the important findings are the decline in noun token proportion and the rise in verb token proportion observed in Murdoch's and Christie's novels (see Figures 8a and 9a). These trends are statistically significant, with *P*-value between 0.0076 and 0.0469 (see Table 12). Our statistical tests also revealed a negative correlation between the noun and verb proportions of the two authors, which is stronger and more significant in Christie's results (see Table 13). These directions of change resemble those found in semantic dementia patients (Bird et al., 2000) in that the apparent noun deficit was compensated for by a rise in verbs. On the other hand, no significant trend is found in James's noun proportion

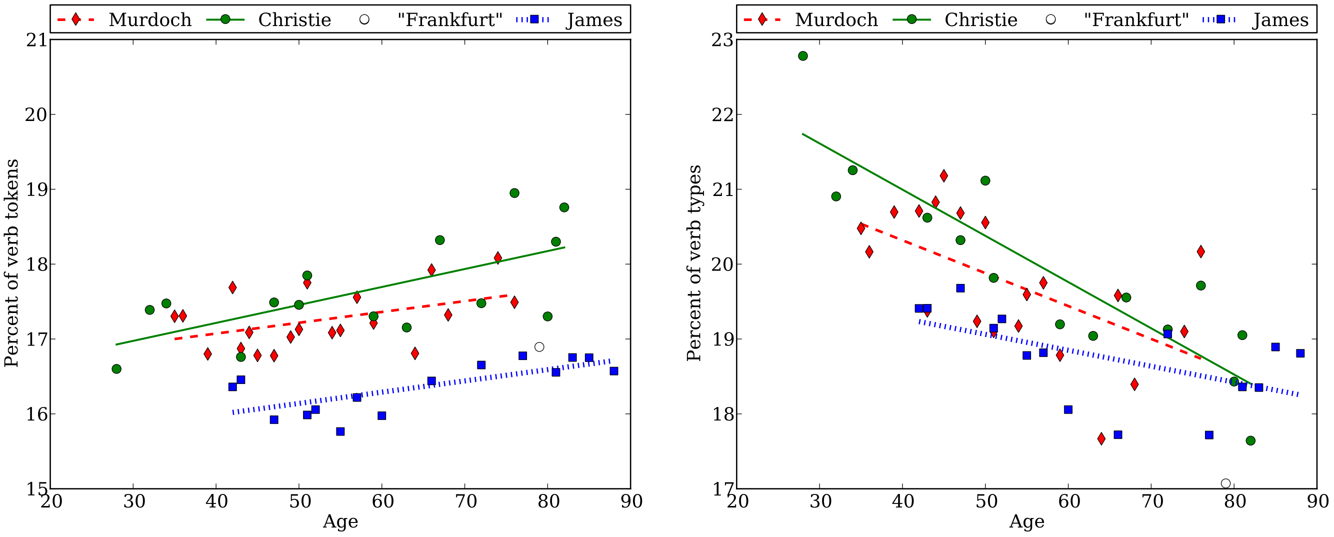


(a) by token

(b) by type

Figure 8: Proportion of common nouns

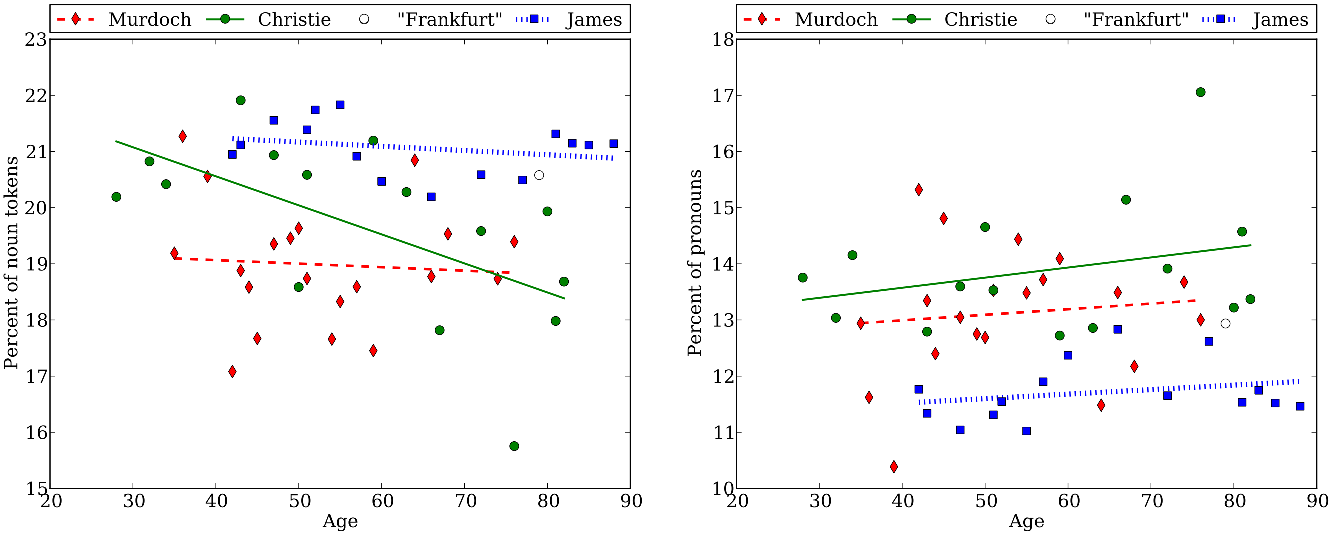
and, although verb proportion shows a slight increasing tendency with high significance, the two values are not highly correlated. Nor are the proportion of pronouns and that of nouns for all three authors, as shown in Table 13.



(a) by token

(b) by type

Figure 9: Proportion of content verbs



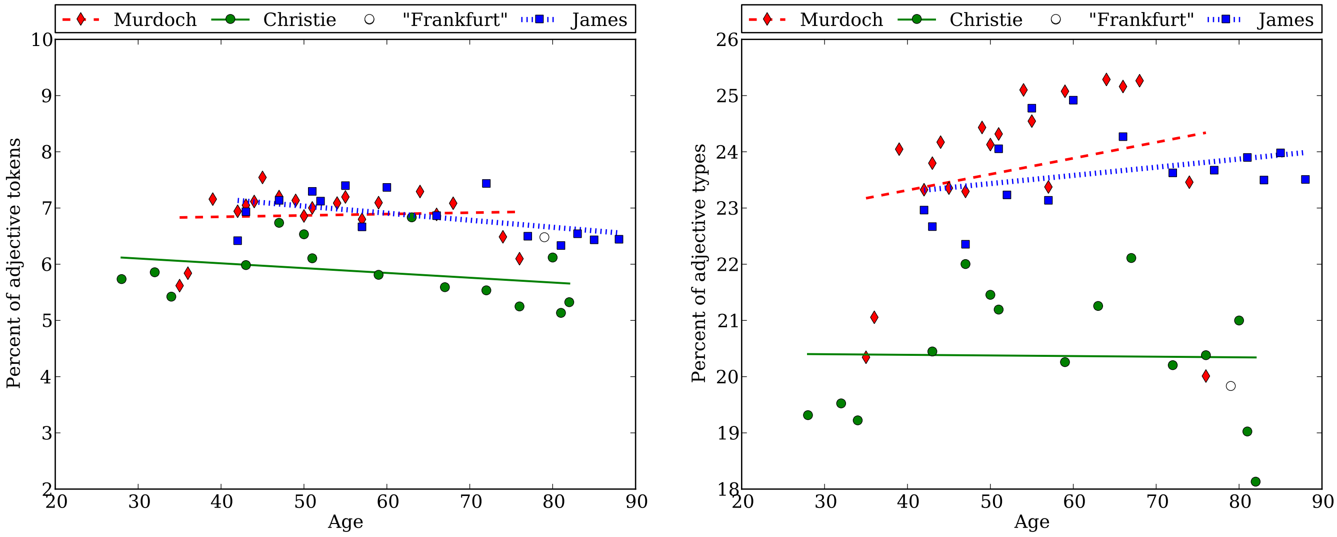
(a) Common and proper nouns

(b) Pronouns

Figure 10: Proportion of nouns and pronouns

However, when proper nouns are considered together with common nouns, a strong negative correlation is found between noun token proportion and pronoun token proportion for all three authors (see Table 13). Figure 10 presents the changes in percentage over time. Again, very few variations exist across James's novels, while Murdoch's and Christie's results span a wide range. As Table 12 shows, a significant decreasing tendency is found in Christie's noun token results, largely due to the sudden drop in her 1967 novel, *Endless Night*. These observations, unsurprisingly, suggest that the deficit in noun is remedied by increased usage of pronouns, in addition to the previously reported rise in verb proportion.

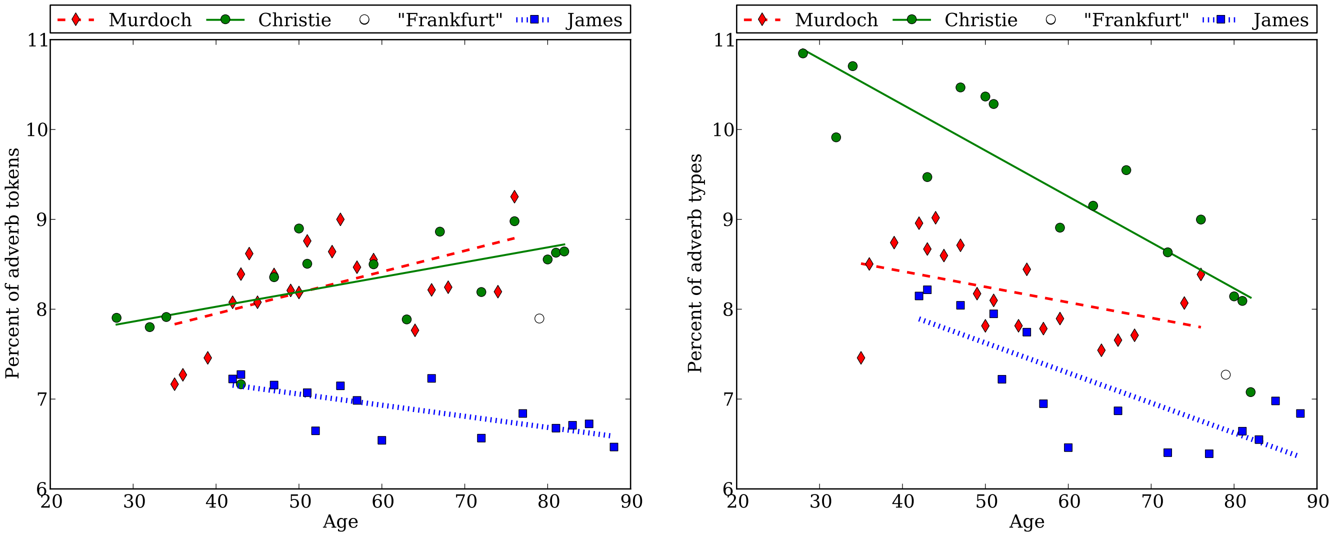
An opposite tendency is observed when types are considered instead of tokens (Figures 8b and 9b). Noun proportions increase while verb proportions decrease for all three authors at varying degrees. These trends are all significant, except for Murdoch's noun proportion (see Table 12). Christie's results have the steepest rate of change and a strong negative correlation between noun and verb, while for Murdoch and James, the change is more gradual and the correlation less pronounced. This fact, combined with the vocabulary and high-frequency verb results, suggests that the decline in Christie's vocabulary is more dramatic for verbs than for nouns, causing an increase in noun type proportion (which does not necessarily signify a growth in noun vocabulary).



(a) by token

(b) by type

Figure 11: Proportion of adjectives



(a) by token

(b) by type

Figure 12: Proportion of adverbs

Similarly, a disconnection between type and token exists in the proportions of adjectives and adverbs (Figures 11–12). While the adjective token proportions remain relatively stable, wide variations are observed in type proportions for all three authors, although none of these trends is statistically significant. An abrupt drop can be seen in Murdoch's and Christie's type proportions in their later novels.

With regard to adverbs, Murdoch's and Christie's token proportions exhibit a statistically significant increase, while James's rates decline slightly. When types are considered, all three authors have a decreasing tendency overall which, as shown in Table 12, is steepest and highly significant for Christie, moderate and significant for James, and slightest and approaching significance for Murdoch.

From the correlation coefficients reported in Table 13, the rise in verb token proportion of Christie's novels is positively correlated with the rise in adverb token proportion, while this is not the case for Murdoch and James. In light of our high-frequency verb results, which reveal that Christie relied heavily on common, less-specific verbs in her later novels, this increased usage of adverbs was perhaps a remedy for the reduced number of specific verbs available in her active vocabulary.

Table 12: Statistical significance test results for word class proportions

	Murdoch		Christie		James	
	Coeff.	$F(1, 18)$	Coeff.	$F(1, 13)$	Coeff.	$F(1, 13)$
Common noun token	–0.0261	7.58^{**}	–0.0295	8.86^{**}	–0.0063	0.41
Common noun type	0.0287	1.83	0.0866	17.25^{**}	0.0336	10.33^{**}
Proper noun token	0.0198	1.18	–0.0222	2.55	–0.0012	0.01
Proper noun type	0.0153	2.91	0.0216	3.44	0.0172	8.98^{*}
Noun token	–0.0062	0.09	–0.0517	7.22^{*}	–0.0075	0.89
Noun type	0.0439	2.44	0.1082	15.32^{**}	0.0507	12.90^{**}
Pronoun token	0.0099	0.19	0.0180	1.23	0.0080	0.81
Content verb token	0.0144	4.55^{*}	0.0240	9.97^{**}	0.0150	13.13^{**}
Content verb type	–0.0439	8.94^{**}	–0.0617	48.84^{**}	–0.0213	5.90^{*}
Adjective token	0.0024	0.06	–0.0085	1.30	–0.0125	4.17
Adjective type	0.0284	0.93	–0.0011	0.00	0.0146	1.53
Adverb token	0.0233	7.09^{*}	0.0165	7.71^{*}	–0.0124	12.28^{**}
Adverb type	–0.0173	4.10	–0.0511	40.48^{**}	–0.0334	21.47^{**}

$^{*}P < 0.05$ $^{**}P < 0.01$

Table 13: Correlation between word class proportions (in token)

		1	2	3	4	5	6	7
MURDOCH	1. Common Noun	—	−0.17	+0.42	-0.53^{*}	+0.18	-0.62^{**}	−0.43
	2. Proper Noun		—	$+0.80^{**}$	+0.06	−0.15	−0.06	-0.66^{**}
	3. Noun			—	−0.19	−0.12	−0.43	-0.88^{**}
	4. Content Verb				—	-0.76^{**}	+0.14	+0.29
	5. Adjective					—	+0.10	−0.04
	6. Adverb						—	+0.33
	7. Pronoun							—
CHRISTIE	1. Common Noun	—	+0.34	$+0.63^{*}$	-0.75^{**}	+0.48	-0.87^{**}	−0.46
	2. Proper Noun		—	$+0.91^{**}$	−0.43	+0.48	-0.56^{*}	-0.75^{**}
	3. Noun			—	-0.59^{*}	+0.51	-0.76^{**}	-0.79^{**}
	4. Content Verb				—	-0.55^{*}	$+0.72^{**}$	$+0.61^{*}$
	5. Adjective					—	−0.32	−0.47
	6. Adverb						—	$+0.64^{**}$
	7. Pronoun							—
JAMES	1. Common Noun	—	-0.55^{*}	+0.29	−0.44	$+0.79^{**}$	−0.39	−0.25
	2. Proper Noun		—	$+0.52^{*}$	−0.05	-0.55^{*}	+0.28	−0.46
	3. Noun			—	−0.39	+0.12	+0.03	-0.82^{**}
	4. Content Verb				—	-0.55^{*}	−0.27	+0.36
	5. Adjective					—	−0.00	−0.24
	6. Adverb						—	−0.13
	7. Pronoun							—

$^{*}P < 0.05$ $^{**}P < 0.01$

Table 14: Correlation between word class proportions (in type)

		1	2	3	4
MURDOCH	1. Common Noun	—	-0.53^{*}	-0.08	-0.64^{**}
	2. Content Verb		—	-0.64^{**}	$+0.60^{**}$
	3. Adjective			—	-0.40
	4. Adverb				—
CHRISTIE	1. Common Noun	—	-0.83^{**}	-0.34	-0.79^{**}
	2. Content Verb		—	$+0.05$	$+0.90^{**}$
	3. Adjective			—	$+0.30$
	4. Adverb				—
JAMES	1. Common Noun	—	-0.76^{**}	$+0.34$	-0.89^{**}
	2. Content Verb		—	-0.67^{**}	$+0.78^{**}$
	3. Adjective			—	-0.43
	4. Adverb				—

$^{*}P < 0.05$ $^{**}P < 0.01$

Lexical marker L5: Fillers

The proportions of lexical fillers and interjections are shown in Figure 13. Consistent with our prediction, Murdoch's and Christie's results indicate clear rising tendencies that are both significant [$F(1,18) = 10.98$, $P = 0.0039$ and $F(1,14) = 6.22$, $P = 0.0258$, respectively]. While the rates of Murdoch's last two novels are only slightly higher than her average results, Christie's rates surge in her last two novels to a peak of 1.67, which is more than double the average of her earlier works (0.79), and nearly triples the lowest rate attained in her 30s (0.55). James's results, on the contrary, remain consistently low throughout, following a slight decreasing trend that is not statistically significant [$F(1,13) = 1.60$, $P = 0.2282$].

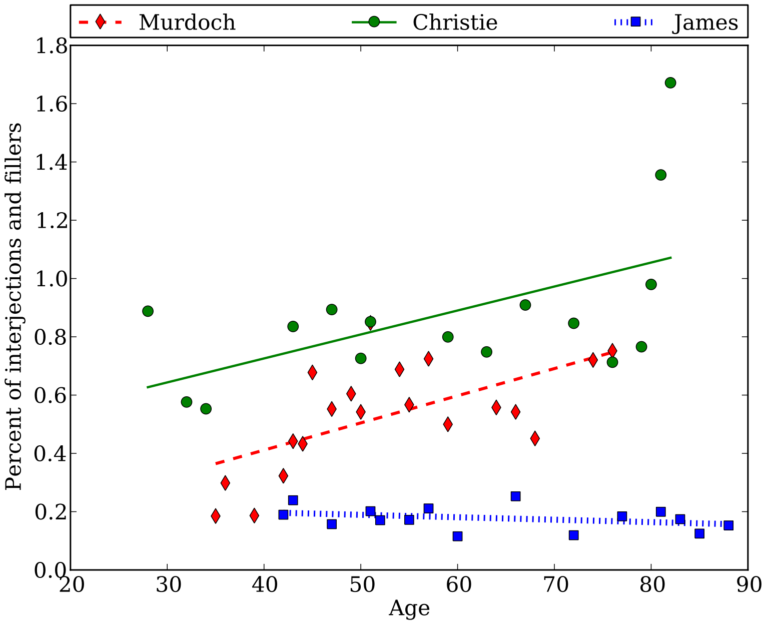


Figure 13: Proportion of interjections and fillers

As Table 15 shows, this measure is moderately correlated, with significance, with other lexical measures for Christie: the correlation is negative for the vocabulary measures—type/token ratio (TTR) and word-type introduction rate (WTIR)—and positive for lexical repetitions (LR) and high-frequency verb proportions (HFV). These results make intuitive sense, because a higher rate of fillers may indicate word-finding difficulty, which leads to a smaller vocabulary size, more repetitions, and greater reliance on generic verbs. For James, a similar correlation is found between this measure and most other measures, with the exception of LR, whereas for Murdoch the situation is reversed. As discussed earlier, the inclusion of fillers in fiction novels may either reflect the speaking style of the writer, or indicate a conscious stylistic choice. Because of the varying degrees of correlation, we neither reject these results nor use them as a basis for our conclusion.

Table 15: Correlation between proportion of fillers and other lexical measures

	TTR	WTIR 50,000	LR	HFV
Murdoch	+0.05	−0.03	$+0.68^{**}$	−0.27
Christie	-0.50^{*}	-0.50^{*}	$+0.60^{*}$	$+0.50^{*}$
James	-0.59^{*}	-0.61^{*}	+0.48	$+0.72^{**}$

$^{*}P < 0.05$ $\quad$ $^{**}P < 0.01$

5.2 Syntactic Results

Syntactic marker S1: Overall syntactic complexity

We now present the results of our syntactic complexity analysis. Statistical tests and correlation coefficients are reported in Tables 16 and 17 (p. 70).

- *Mean number of Clauses per Utterance (MCU):*

The MCU results of each author suggests an overall increasing trend, which is significant for Christie and James. Christie's syntax is relatively lower in number of clauses, which fluctuate between 1.65 and 1.93 before rising in her last few novels to a maximum of 2.13. James's results also hint at an upward trend, varying between 2.07 and 2.45, and peaking at her last novel. Murdoch's overall trend is insignificant because of a deep drop around her late-40s and 50s, of which the lowest point is at the 51-year mark with her 1970 novel, *A Fairly Honorable Defeat*.

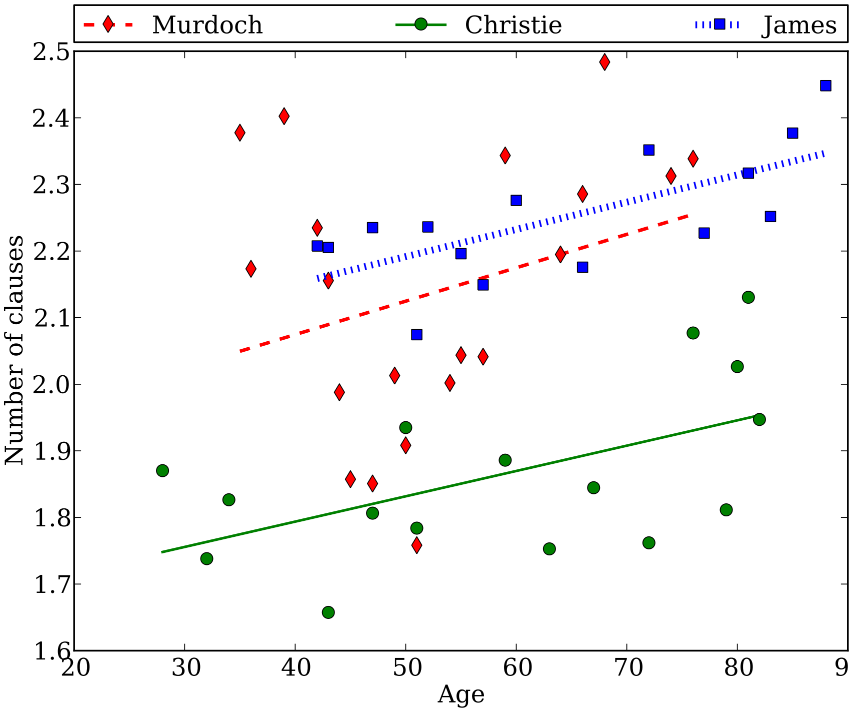


Figure 14: Mean number of clauses per sentence

- *Mean Length in words per Utterance (MLU):*

The MLU results of all three authors again show increasing tendencies, none of which are statistically significant. Murdoch's mean sentence length reaches a peak early in her career, then plummets to a low at age 51 (which coincides with Christie's datapoint at 51 in Figure 15), and gradually recovers in her later works before dropping slightly with her final two novels. On the other hand, Christie's sentence length stays relatively stable before climbing to a peak at age 80, then declines to her usual range with her last novel. James's mean sentence length fluctuates between 13.18 and 14.76, and reaches its peak at her latest work.

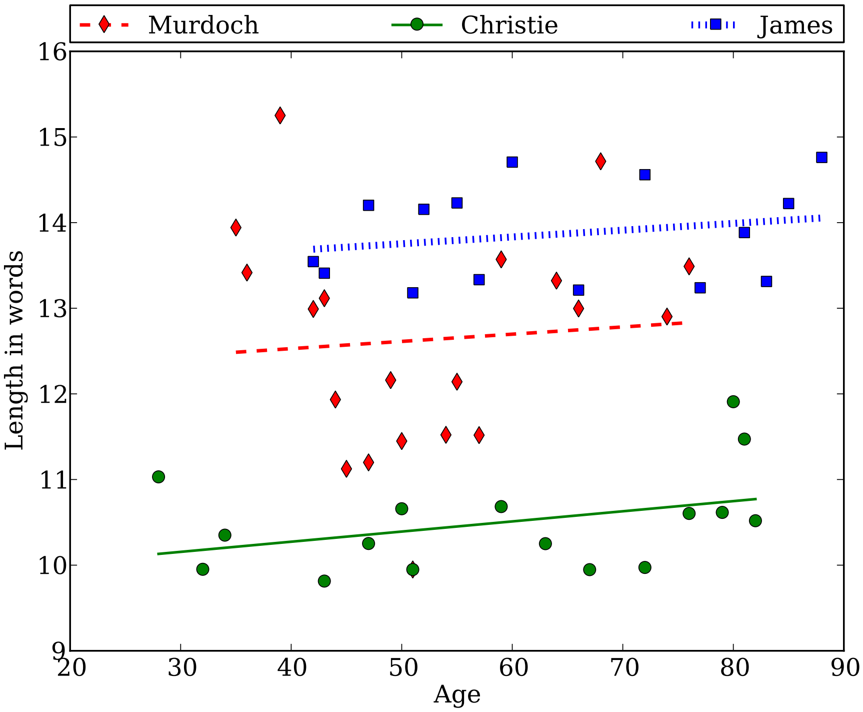


Figure 15: Mean length in words per sentence

- *Average parse tree depths:*

Figure 16 displays the average unweighted parse tree depth of each novel. Murdoch's results follow a pattern similar to those of her MCU and MLU results: a brief rise in the early novels, a steep drop in her 40s and 50s, followed by a period of recovery and a drop in her last novels; the overall trend is a statistically insignificant decrease. In contrast, the parse tree depths of Christie's novels constitute a significant increase, which is in part due to the sharp rise in her later novels. James's average depth remains consistent throughout her career and increases slightly in her two most recent novels; the overall increasing trend is insignificant.

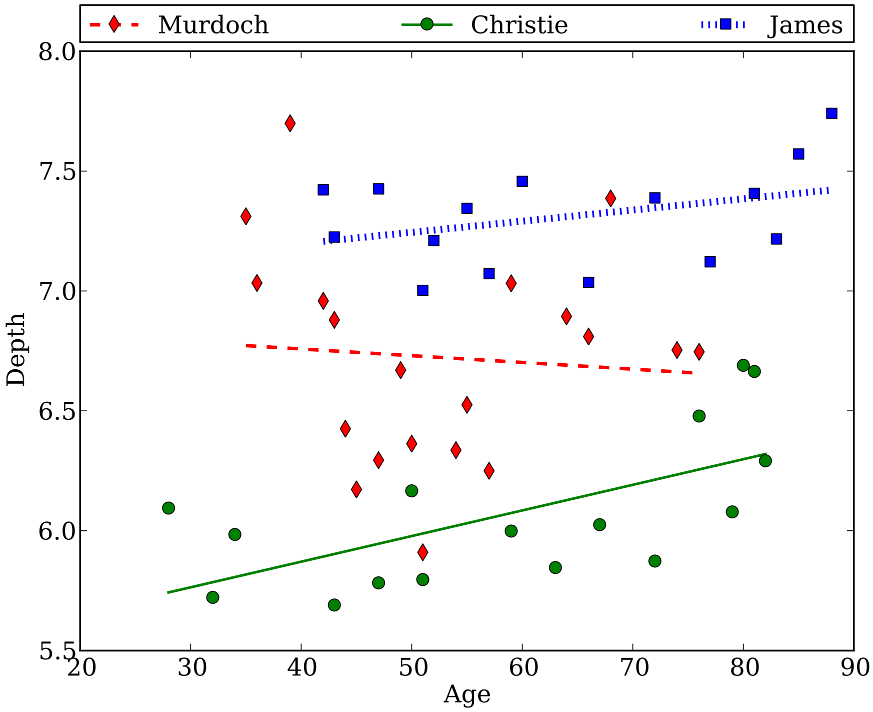
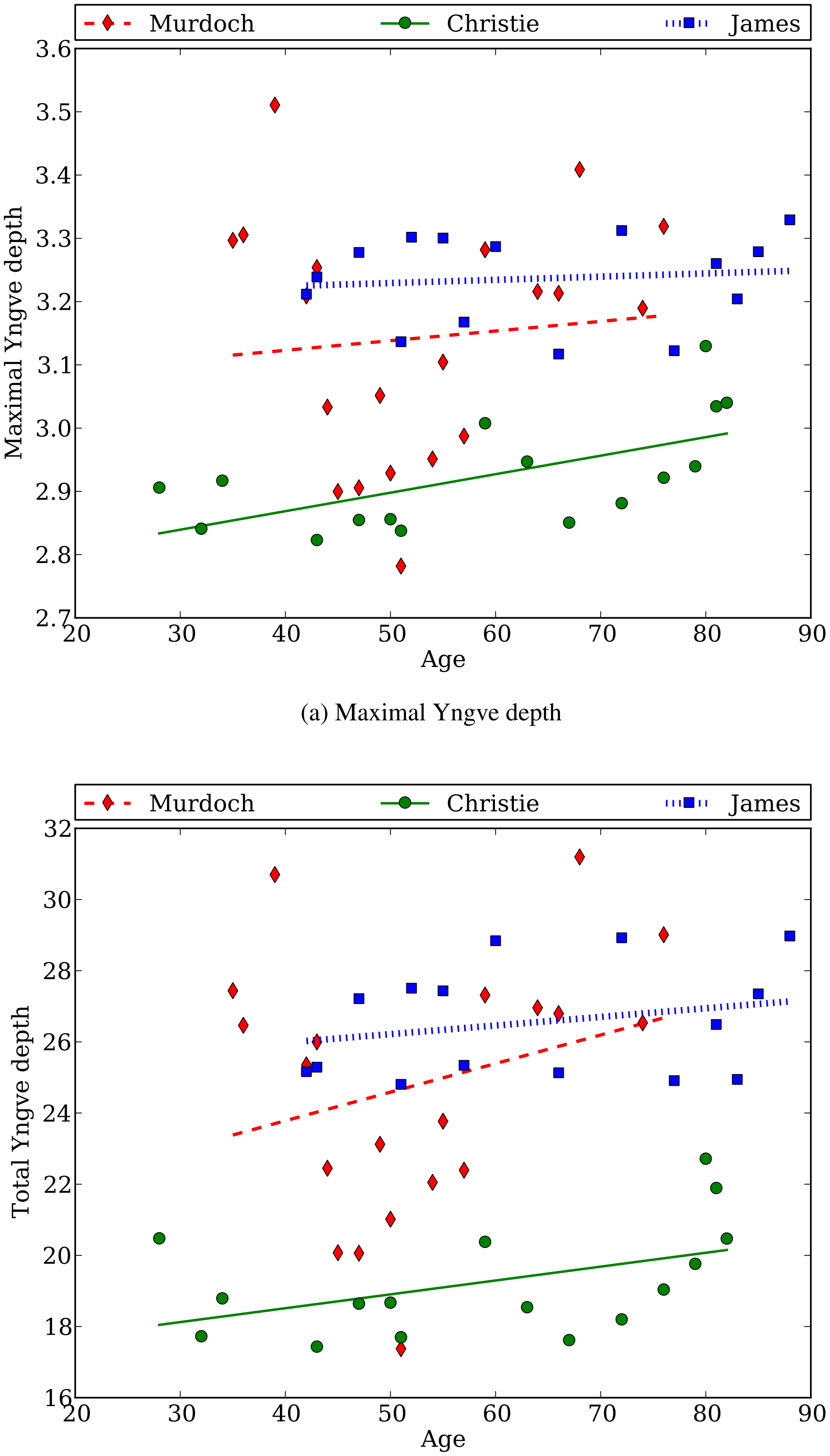


Figure 16: Unweighted parse tree depth

The results of Yngve depth measures, which assign more weight to left-branching structures, are shown in Figure 17. A pattern resembling those of the previous syntactic measures is found in Murdoch's novels, both in terms of maximal and total Yngve depths; the overall linear trend is again insignificant. Statistical tests on James data yield insignificant results for the rise in both measures, while the maximal depths of Christie's novels show a small but significant increase.



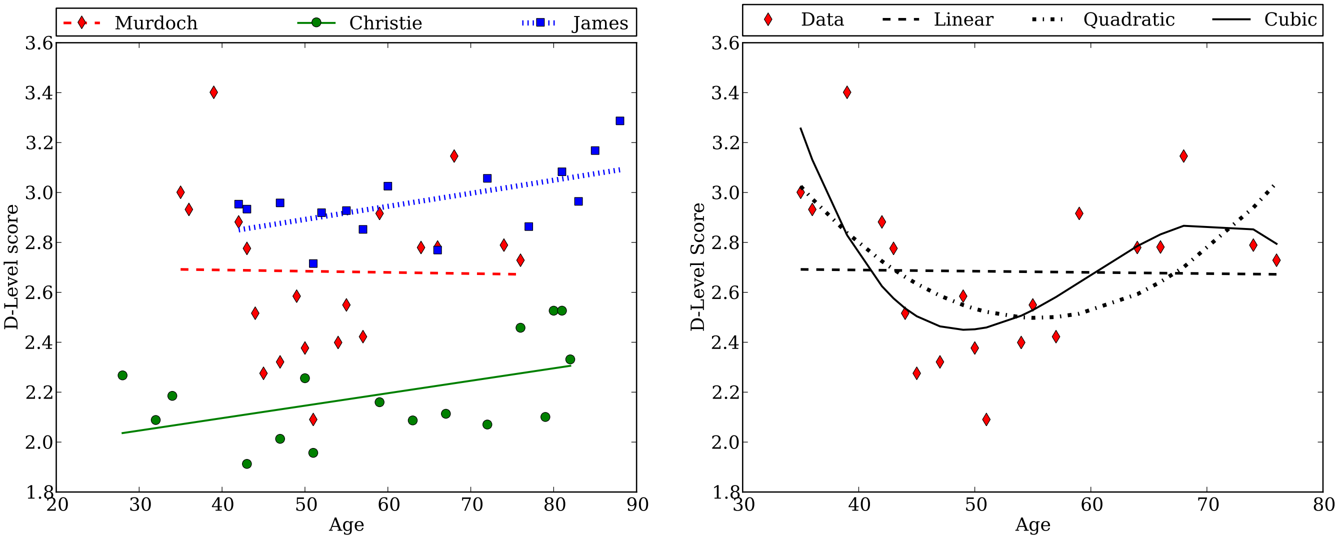
(a) Maximal Yngve depth

(b) Total Yngve depth

Figure 17: Average Yngve depths

- *D-Level score:*

Figure 18a displays results of the D-Level measure, which are largely similar to those of other syntactic measures. Murdoch's average D-Level scores over time exhibit a very slight decrease, which is statistically insignificant. In contrast, statistical tests indicate an upward trend that approaches significance for Christie [$F(1,14) = 4.46$, $P = 0.0531$], and one that is highly significant for James [$F(1,13) = 6.33$, $P = 0.0258$].



(a) All three authors

(b) Murdoch

Figure 18: Average D-Level score

As summarized in Table 16, few of the syntactic complexity measures yield statistically significant results. This reflects the lack of linear rising or falling trends in the data; Murdoch's D-Level results, for instance, are best represented by a cubic regression model, compared to the linear and quadratic counterparts, as demonstrated in Figure 18b.

Table 17 shows the correlation coefficients between the complexity measures, which are mostly moderate to high, especially for Murdoch. The high level of agreement among different measures reconfirms that, compared to her earlier works, Murdoch's syntactic complexity undergoes a period of relatively steep decrease around the author's late-40s and 50s, followed by a period of gradual increase evident in her later novels.

Table 16: Statistical significance test results of syntactic complexity measures

	Murdoch		Christie		James	
	Coeff.	$F(1, 18)$	Coeff.	$F(1, 14)$	Coeff.	$F(1, 13)$
MCU	0.0050	1.66	0.0038	6.08*	0.0041	11.91**
MLU	0.0084	0.11	0.0118	2.37	0.0079	0.66
Unw.Depth	–0.0028	0.10	0.0107	9.44**	0.0047	1.91
Max.Yngve	0.0015	0.17	0.0029	8.70*	0.0005	0.16
TotalYngve	0.0803	1.33	0.0390	3.88	0.0242	0.83
D-Level	–0.0005	0.01	0.0050	4.46	0.0052	6.33*

$^{*}P < 0.05$ $^{**}P < 0.01$

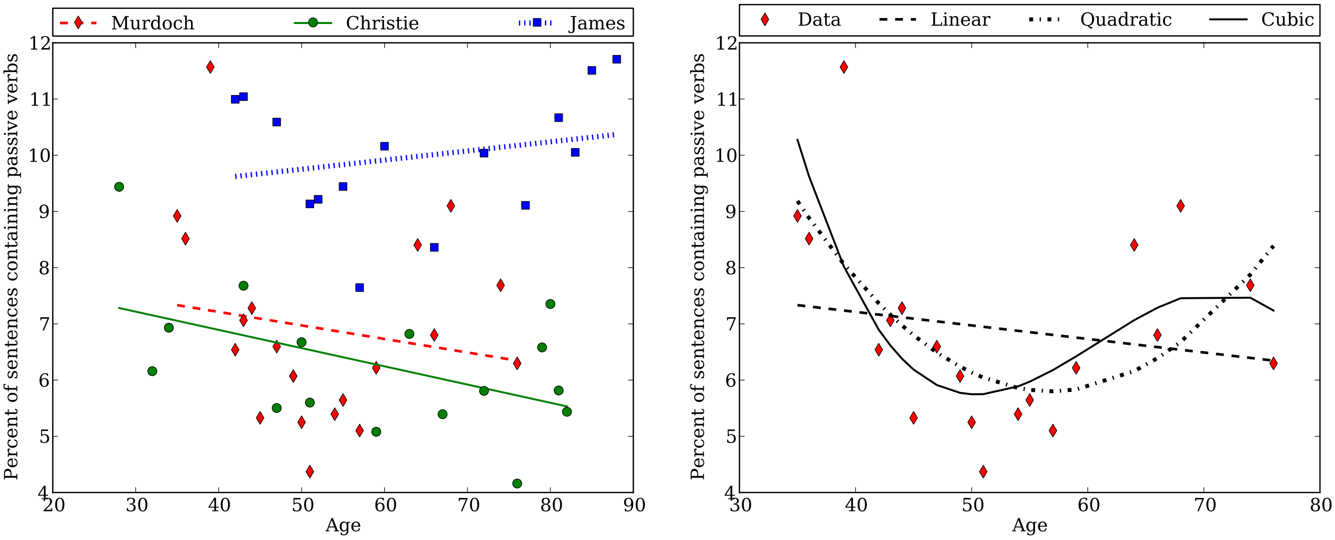
Table 17: Correlation between syntactic complexity measures

		1	2	3	4	5	6
Murdoch	1. MCU	—	+0.94**	+0.88**	+0.92**	+0.97**	+0.93**
	2. MLU		—	+0.95**	+0.98**	+0.97**	+0.94**
	3. Unw.Depth			—	+0.92**	+0.89**	+0.98**
	4. Max.Yngve				—	+0.96**	+0.91**
	5. Total Yngve					—	+0.90**
	6. D-Level						—
Christie	1. MCU	—	+0.78**	+0.94**	+0.67**	+0.78**	+0.92**
	2. MLU		—	+0.79**	+0.76**	+0.95**	+0.80**
	3. Unw.Depth			—	+0.74**	+0.80**	+0.93**
	4. Max.Yngve				—	+0.86**	+0.73**
	5. Total Yngve					—	+0.83**
	6. D-Level						—
James	1. MCU	—	+0.70**	+0.77**	+0.68**	+0.65**	+0.93**
	2. MLU		—	+0.84**	+0.93**	+0.95**	+0.75**
	3. Unw.Depth			—	+0.69**	+0.68**	+0.89**
	4. Max.Yngve				—	+0.94**	+0.67**
	5. Total Yngve					—	+0.63*
	6. D-Level						—

$^{*}P < 0.05$ $^{**}P < 0.01$

Syntactic marker S2: Passive voice

The passive structures detected by our program include explicit passives containing the verb *be* or *get*, and bare passives preceding a *by*-phrase. Figure 19a shows the proportion of sentences containing these passive forms over the total number of sentences in each text. James's results indicate a slight upward trend, while Murdoch's and Christie's exhibit a decline. None of these trends is statistically significant, as summarized in Table 18; Christie's decline, however, approaches significance with a P-value of 0.0541.



(a) All three authors

(b) Murdoch

Figure 19: Proportion of passive sentences

Table 19 shows that this measure is moderately correlated with most syntactic complexity measures for Murdoch and James, but this is clearly not the case for Christie. These facts suggest that access to passive forms may be, though not necessarily, affected by the overall complexity of one's syntax. Similar to the complexity results, Murdoch's passive proportion is best modelled by cubic regression, as demonstrated by Figure 19b, which is consistent with the previous observation of a syntactic decline in her 50s.

Table 18: Statistical significance test results of passive voice measures

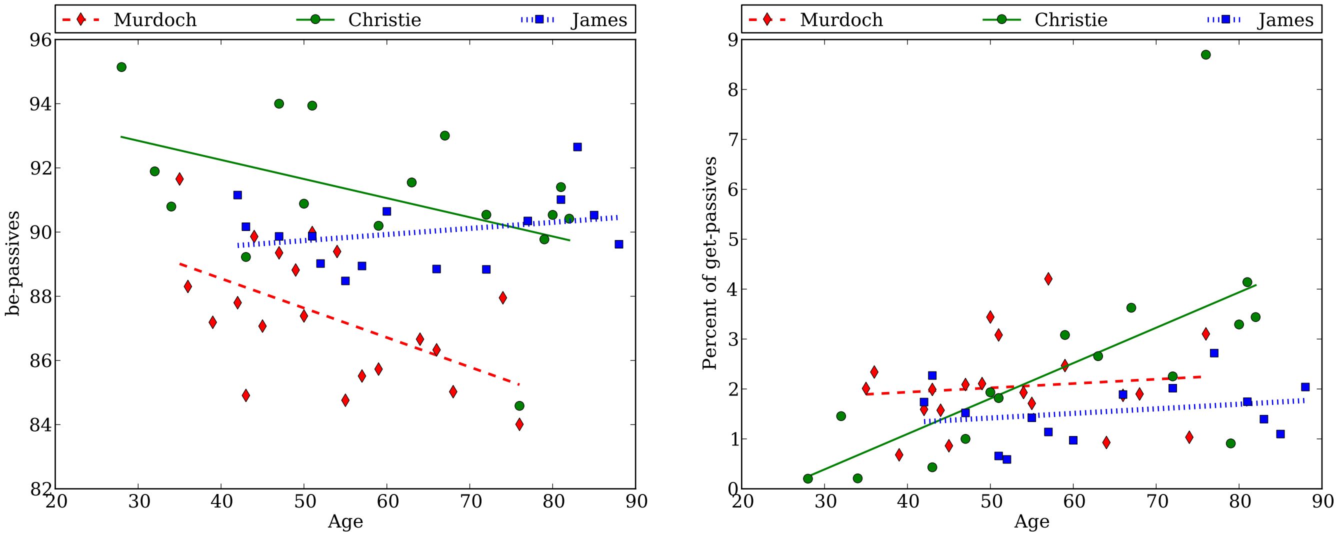
	MURDOCH		CHRISTIE		JAMES	
	Coeff.	$F(1, 18)$	Coeff.	$F(1, 14)$	Coeff.	$F(1, 13)$
Passive sentences	–0.0241	0.53	–0.0324	4.42	0.0162	0.69
Sentences with *be*-passives	–0.0918	7.42^{*}	–0.0596	3.82	0.0190	1.03
Sentences with *get*-passives	0.0086	0.25	0.0709	9.43^{**}	0.0092	0.84
Sentences with *by*-phrase	0.1573	8.86^{**}	–0.0671	3.35	–0.0189	0.34

$^{*}P < 0.05$ $^{**}P < 0.01$

Table 19: Correlation between passive proportion and syntactic complexity measures

	MCU	MLU	Unw.Depth	Max.Yngve	Total Yngve	D-Level
Murdoch	$+0.69^{**}$	$+0.77^{**}$	$+0.83^{**}$	$+0.70^{**}$	$+0.69^{**}$	$+0.81^{**}$
Christie	−0.27	+0.14	−0.09	+0.19	+0.06	−0.04
James	$+0.65^{**}$	$+0.56^{*}$	$+0.85^{**}$	+0.20	+0.18	$+0.81^{**}$

$^{*}P < 0.05$ $^{**}P < 0.01$



(a) *be*-passives

(b) *get*-passives

Figure 20: Proportions of *be*-passives and *get*-passives and

Figures 20a and 20b show the proportion of passive sentences that contain the verbs *be* and *get*, respectively. The proportions of *be*-passives in Murdoch's and Christie's novels both exhibit a declining trend, which is stronger and significant for Murdoch. James's *be*-passives, on the other hand, increase in proportion, though without significance. With respect to *get*-passives, a mild increase exists in Murdoch's and James's results; Christie's results, in contrast, follow a strong, highly significant rising pattern, climbing to a peak abruptly at her 1967 novel, *Endless Night*. Proportions of passives with *by*-phrase, shown in Figure 21, suggest a moderate decline over time for Christie, a very slight decline for James and, surprisingly, a significant increase for Murdoch. Statistical test results for these measures can be found in Table 18.

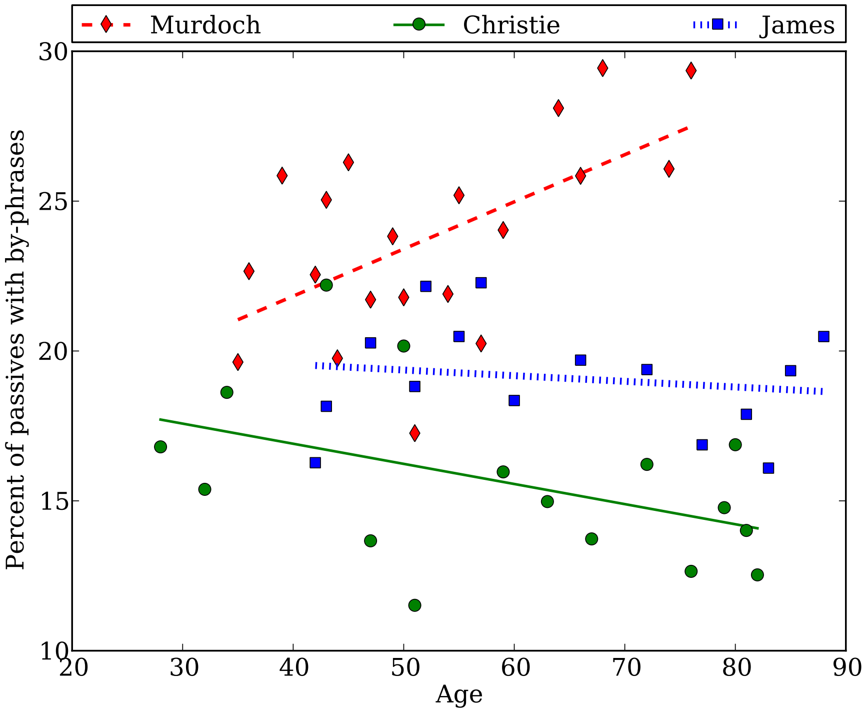


Figure 21: Proportion of passive sentences with *by*-phrase

5.3 Discussion

Lexical Analysis:

Our lexical analysis yields results that largely follow our hypothesis and the expected patterns of linguistic change given in Table 1 (p. 15). Murdoch's lexical decline is evident in the type/token ratios and word-type introduction rates of her later novels, especially *Jackson's Dilemma*, which shows an abrupt decline in vocabulary size in the latter half of the book. As expected, the decline in vocabulary leads to a significant increase in lexical repetitions of content words, and a word class deficit can be seen in noun token proportion, with compensation in verb token proportion. Contrary to our prediction, Murdoch's verb specificity appears intact.

The lexical results of P. D. James's novels follow the predicted patterns for normal aging elders. Her vocabulary size, lexical repetition, and verb specificity vary in a relatively small range, with no apparent word class deficit. The word-type introduction measure reveals a slight tapering in the vocabulary growth of the latter half of James's latest book, *The Private Patient*; however, unlike Murdoch's, the overall rate does not stray far from James's average range, and a mild age-related decline is expected of the vocabularies of healthy elders.

Because of the change in her writing technology in 1952, Christie's lexical analysis yields interesting results, showing an overall decline and, in addition, an effect of revisions, research, and editorial help. Our vocabulary, repetition, and verb specificity measures reveal a relatively consistent linguistic decline, not only since Christie started using the dictaphone in 1952, but since the start of her career. *Destination Unknown*, published in 1954 when she was 63, contains a larger vocabulary, fewer lexical repetitions, and a lower proportion of high-frequency verbs, quite distinct from the other novels composed in the same period. The differences resemble an effect of revisions, which Christie acknowledged, in her autobiography, to have become necessary with the use of the dictaphone, and which she also deemed "irritating."[1] The obvious decline in her subsequent novels suggests either an increasingly lax revision process, or that the age- or disease-related effect gradually became more prevalent. A similar situation is observed with *Passenger to*

[1] Agatha Christie. *Agatha Christie: An Autobiography*. Dodd, Mead and Company, 1977: 348.

Frankfurt. Not only does the novel exhibit the largest vocabulary out of the sixteen Christie novels we analyzed, *Frankfurt* also has the highest word-type introduction rate, and significantly fewer lexical repetitions and high-frequency verbs, compared to the other novels written in Christie's late-70s and early 80s. Her final novel, *Postern of Fate*, which involved editorial help, registers a small but noticeable improvement in vocabulary, as well as a smaller number of repetitions and high-frequency verbs, compared to her penultimate work, *Elephants Can Remember*. In addition, consistent with our prediction, a deficit in noun tokens is found, which is significantly correlated with the rise in verb and pronoun tokens; however, when types are considered, the decline in verbs appears to be more dramatic than in nouns. Christie's adverb token proportion also increases with significance and, since it is highly correlated with the rise in verbs, may have been a remedy for the loss of specific, descriptive verbs.

One lexical measure contradicts our predicted patterns of changes, namely, noun specificity, approximated using synset depths in WordNet noun hierarchy. The results for Christie and James are statistically insignificant, whereas for Murdoch, both noun type and noun token results indicate a slight but significant increase. However, the overall irregularity of the results, combined with the inconsistencies between types and tokens, led us to question the accuracy of this approximation. We address the limitations of this measure in section 6.2, and the fitness of depths in WordNet as an approximation of specificity in Appendix B.2.

With the exclusion of the noun specificity approximation for all authors, and the high-frequency verb measure for Murdoch, our lexical analysis discovers the linguistic patterns typical of normal aging adults in the novels of P. D. James, and the patterns of decline observed in dementia patients in Iris Murdoch's and Agatha Christie's writings. The rates of change and the linguistic behaviours are not identical between Murdoch's and Christie's results, a testament to the heterogeneity in linguistic changes, asserted by Maxim and Bryan (1994), among individuals in both normal aging and dementia. However, Murdoch's and Christie's works exhibit sufficiently similar trends—and decidedly distinct from the patterns found in James's novels—to validate our hypothesis with respect to lexical linguistic markers.

Syntactic Analysis:

In contrast to our lexical results, little can be said with certainty about the syntactic results, because of the lack of significant linear trends over each entire dataset. In particular, no significant linear trends are found in Murdoch's novels for the entire period, but all syntactic measures consistently reveal an abrupt drop in her late 40s and 50s, then a period of recovery which, for some measures, is followed by a slight decline in her last two novels. This early syntactic decline is not a singular occurrence in a few novels, but all of the novels in that period (this rules out the possibility that some of the digitized texts contain a larger number of data errors than others). While this pattern in Murdoch's syntax is seemingly unconnected with her lexical results, a mild drop in type/token ratio, the peak in lexical repetitions, and an increase in high-frequency verb proportion can be detected in the novels written in her early 50s. Our theory to explain this phenomenon is that either Murdoch was experimenting with a different writing style, or the early decline signifies the pathology of AD (which, as mentioned earlier, is insidious and may begin many years or decades before the disease onset). However, the former explanation only applies to syntax and does not justify the lexical decline, and the latter does not account for the gradual recovery that followed.

Similar to her lexical results, most of James's syntactic results vary only slightly, with the widest span being the passive sentence proportion. Christie's results fluctuate in a relatively wider range. The overall trends for both authors indicate a rising tendency in all measures, although only a few yield significant results. This stands in sharp contrast with the marked decline in lexical features observed in Christie's novels. If the author indeed had dementia, these facts support the widely held belief that syntax is relatively spared, and that a core grammatical system is still preserved and functional, even in severe cases of dementia (Maxim and Bryan, 1994). On the other hand, Bates et al. (1995) were clear in their premise that the observed deficit in passive structure production only emerges in highly constrained situations which present a natural context for passive sentences—the participants in their study rarely produced any passives in the free description task. Novels provide a similar setting in which no such constraints are imposed. While James's results indicate an insignificant increase in passive proportion, Murdoch's and Christie's passive results follow the same direction that Bates et al. (1995) documented (a decrease in the passive sentence proportion, a rise in *get*-passives, and a drop in *be*-passives), although, without constraints, only Murdoch's *be*-passive and Christie's *get*-passive results are significant. A result that contradicts the expected passive patterns is Murdoch's proportions of passives with *by*-phrase, which increase with high significance.

Overall, the syntactic analysis yields results that are less definite than our lexical analysis does. A decline is found in Murdoch's novels, following a cubic model rather than a linear one. James's syntactic results follow the patterns expected of healthy elders; however, our hypothesis that Christie's patterns of changes resemble Murdoch's does not hold.

Table 20 summarizes the changes observed in the novels of Murdoch, Christie, and James, with respect to the patterns of linguistic changes reported by other studies of language in normal aging and dementia, given in Table 1 (p. 15). The items reported in parentheses indicate statistically insignificant trends. Check marks indicate that the patterns observed follow our hypotheses; crosses indicate otherwise.

Table 20: Patterns of linguistic changes observed in the novels of Murdoch, Christie, and James

		LINGUISTIC MARKER	MURDOCH	CHRISTIE	JAMES
		Lexical:			
✓	(L1)	Vocabulary size:	sharp decrease in last novel; signs of decline in her 50s	gradual decrease overall; sharp decrease in later novels	(gradual increase overall); (marginal decrease in later novels)
✓	(L2)	Lexical repetition:	strong increase overall; sharp rise in her 50s	pronounced increase	(marginal increase)
×	(L3)	High-frequency verb proportion:[1]	moderate decrease; noticeable rise in her 50s	pronounced increase in proportion	(slight decrease)
		Noun specificity:[2]	slight increase in tokens and types	(decrease in tokens); (increase in types)	(decrease in tokens); (increase in types)
✓	(L4)	Word class deficit:	deficit in noun tokens; compensation in verb tokens	deficit in noun tokens; compensation in verb tokens	(marginal decrease in noun tokens); (uncorrelated rise in verb tokens)
✓	(L5)	Fillers:[3]	pronounced increase overall; noticeable rise in her 50s	pronounced increase	(slight decrease)
		Syntactic:			
×	(S1)	Overall complexity:	irregular changes; deep decline in her 50s	(minor changes)	(minor changes)
×	(S2)	Use of passive:			
✓		Overall:	(decrease); sharp drop in her 50s	(decrease)	(increase)
✓		*be*-passives:	decrease	(decrease)	(increase)
×		*get*-passives:	(increase); sharp rise in her 50s	increase	(increase)
×		with *by*-phrase:	increase; sharp drop in her 50s	(decrease)	(decrease)

[1] Approximates word specificity for verbs. A higher proportion of high-frequency, generic verbs means a smaller proportion of more specific ones.

[2] Potentially inaccurate approximation.

[3] May also reflect an author's stylistic choices in creating natural dialogues.

6 Conclusion

6.1 Summary of Results

In this study, we conducted lexical and syntactic analyses on fifty-one novels by Iris Murdoch, Agatha Christie and P. D. James. The lexical analysis discovers strong evidence of a linguistic decline in both Murdoch's and Christie's later works, whereas James's results remain relatively stable throughout her career. The lexical measures found to produce the clearest results, and therefore proposed to be sensitive to the effects of dementia on language production, are the two vocabulary measures—namely, type/token ratio and word-type introduction rate—and the measure of repetitions of content words within close distance.

The syntactic analysis largely yields statistically insignificant results. All syntactic measures register a consistent cubic pattern of change in Murdoch's novels; a period of deep decline occurs in her late 40s and early 50s, which coincides with a linguistic decline found by some of the lexical measures. Contrary to our hypotheses, both Christie's and James's syntactic complexity results exhibit a slight *rising* tendency, although few measures discover a significant linear trend. Murdoch's and Christie's use of passives somewhat resembles the patterns observed in AD patients by Bates et al. (1995), but some patterns occur without statistical significance.

The results of our study, summarized in Table 20, provide further support for the hypothesis that signs of dementia can be detected in diachronic analysis of patients' writings, most evidently in the lexical features, and that it is possible to distinguish this disease-related decline from the normal effects of aging.

6.2 Limitations

While we tried to maintain a high accuracy, our approach suffers from limitations leading to possible errors at the various stages, which we have documented below so that future improvements can be made.

Data:

Errors in data are either typographical errors in the source books, which are rare but nonetheless do occur, or OCR errors that have not been caught in our data correction stage. These errors affect the accuracy of all measures to varying degrees.

Sentence boundaries:

Incorrect determination of sentence boundaries is caused by missing or incorrect punctuation in the data, or stylistic differences in punctuation usage. For instance, dashes are sometimes used as sentence-ending markers in Murdoch's novels, whereas Christie tended to use ellipses for the same purpose. Our algorithm, being deterministic and heuristics-based, might not correctly handle uncommon punctuation usage. Errors made at this stage affect measures that operate on a per-sentence basis, more specifically, the syntactic complexity measures.

Syntactic parse trees:

The Charniak parser may build parse trees that contain either incorrect part-of-speech tags, or wrong embedding levels for linguistic components. Example errors of the former type are: tagging all instances of *be*, *have*, or *do* as auxiliaries; and tagging past participles (VBN) in explicit passive structures as preterits (VBD). Errors in embedding levels often occur for structures that involve possessive determiner modifying gerund; for instance, Level 3 sentence example (given by Covington et al., 2006) for nominalization in object position is:

> Why can't you understand his rejection of the offer?

The underlined segment is correctly parsed as

```
(NP (NP (PRP$ his) (NN rejection)) (PP (IN of) (NP (DT the) (NN offer))))
```

However, when the gerund *rejecting* is used instead, the resulting parse tree segment becomes

`(NP (PRP his)) (S (VP (VBG rejecting) (NP (DT the) (NN offer)))),` or

`(NP (NNP John) (POS 's)) (S (VP (VBG rejecting) (NP (DT the) (NN offer))))`

In both cases, the possessive determiners *his* and *John's* are taken to be separate noun phrases, not as modifiers of the gerund *rejecting*, which is recognized as a clause marked by the tag *S*.

Some of these parser errors have been corrected; some accounted for by the pattern sets developed for the D-Level and passive measures. Unhandled errors affect measures that are based on parse trees, including all syntactic measures with the exception of mean sentence length, and word class proportion results, which rely on the part-of-speech tags of the parse trees.

In addition, the structure of the Charniak parser output may cause inflation in our tree depth measures (unweighted depth, Yngve maximal depth, and Yngve total depth). For example, the verb phrase *could have been* is parsed as `(VP (MD could) (VP (AUX have) (VP (VBN been))))`, with three sub-levels, rather than the flatter structure, `(VP (MD could) (AUX have) (VBN been))`, with only one sub-level. As a result, sentences containing these types of structure may receive a depth score comparable to or higher than other sentences that are syntactically more complicated, such as those with relative clauses or appositions.

Pattern sets:

The patterns developed for our D-Level and passive voice measures (included in Appendix B.3) may assign incorrect scores to sentences. First, the accuracy of pattern matching depends heavily on the correctness of the parse trees; while our patterns can detect some common parser errors, the program can only fix errors that can be unambiguously identified. Second, the pattern sets are not comprehensive, since several sentence structures cannot be distinguished from others judging from syntax alone. We excluded highly ambiguous structures from the pattern sets, to avoid creating false positives, at the cost of allowing some false negatives; for example, bare passives without *by*-phrase are syntactically identical to adjectival or perfect uses of past participles. On the other hand, defined patterns may also match a relatively small number of false positives. One such example is nominalization, defined by Covington et al. (2006) as "sentences converted into abstract noun phrases, such as *the enemy's destruction of the city* (from *the enemy destroyed the city*)" (5). The pattern for this type of nominalization requires a determiner and a prepositional phrase that modifies a head noun that ends in one of several possible suffixes of nominalizations. However, this pattern also accepts the noun phrase *the philosopher's definition of the mind*, in which *definition* refers to the statement that defines the mind, rather than the act of defining it, and is thus not a case of nominalization.

Word-sense disambiguation and WordNet-based specificity approximation:

Word-sense disambiguation is performed on a per-sentence basis by the WordNet::SenseRelate program (Pedersen, 2009). Errors in determining the correct synsets may occur for sentences in which there are few content words, and sentences whose contexts depend heavily on nearby sentences. Even when the synsets are correctly identified, WordNet depth might be unsuitable as a source for word specificity approximation, because of its uneven degrees of distinction among its noun branches. (A close examination of this issue can be found in Appendix B.2.) The error rates of this measure cannot be assessed automatically—it is difficult even to manually disambiguate word senses with respect to the (sometimes overtly fine-grained) distinctions in WordNet. However, we suspect that the combined error rates of sense disambiguation and specificity approximation might make the accuracy of this measure less than desirable. An obvious direction for improvement is to extend the range of sentences considered in the disambiguation process, and alter the way depth in WordNet is measured, or employ a new approximation technique altogether.

6.3 Future Directions

More data for cross-sectional and longitudinal analyses:

Because the manifestation of dementia may differ among individuals, depending on the type of dementia and the stage of the disease, a larger number of writing samples—including the more informal, spontaneous kinds of writing, such as blogs and emails—by different subjects is needed to discover the general linguistic patterns, if they exist. We also aim to digitize the remaining nine novels of Murdoch and James for textual analysis of the authors' complete bibliographies, and extend our collection of Christie's novels to better represent her writing career.

Inclusion of semantics into analysis:

Semantic analysis, including the study of argument structure and discourse analysis, will shed new light on the linguistic changes in patients compared to those in non-patients. Syntactic analysis will also benefit from the inclusion of semantics in selecting among possible syntactic structures for ambiguous sentences.

Separate analysis of dialogues and narratives:

The current approach does not differentiate between conscious syntactic or lexical changes often found in the characterization process of fiction novels and unconscious changes due to age- or disease-related decline. Once the digitized texts are formatted in such a way that allows accurate detection of dialogues, separate analyses will be performed on the narrative and dialogue portions of the novels. The narrative-only analysis arguably presents a more accurate assessment of the writers' linguistic levels,[1] and a comparison between narrative and dialogue may reveal the impact of stylistic choices on the outcome of our evaluation techniques.

[1] The inclusion of dialogues in the analysis will considerably reduce the average syntactic complexity scores of texts, because of the higher numbers of fragments, false starts, sentences without subjects, sentences that are single-word interjections, or utterances interrupted in mid-sentence by other characters, which are often found in dialogues. These types of syntactic structures, illustrated in the following example, arguably reflect the nature of dialogue more than the syntactic level of the author.

"It's bad news, Flora," he said quietly. "Bad news for all of us. Your Uncle Roger—"

"Yes?"

"It will be a shock to you. Bound to be. Poor Roger's dead." [C3]

Appendices

A Lists of Novels

IRIS MURDOCH

I.D.	YEAR OF PUBLICATION	APPROX. AGE AT COMPOSITION	NOVEL
M1	1954	35	Under the Net
M2	1955	36	The Flight from the Enchanter
M3	1958	39	The Bell
M4	1961	42	A Severed Head
M5	1962	43	An Unofficial Rose
M6	1963	44	The Unicorn
M7	1964	45	The Italian Girl
M8	1966	47	The Time of the Angels
M9	1968	49	The Nice and the Good
M10	1969	50	Bruno's Dream
M11	1970	51	A Fairly Honorable Defeat
M12	1973	54	The Black Prince
M13	1974	55	The Sacred and Profane Love Machine
M14	1976	57	Henry and Cato
M15	1978	59	The Sea, the Sea
M16	1983	64	The Philosopher's Pupil
M17	1985	66	The Good Apprentice
M18	1987	68	The Book and the Brotherhood
M19	1993	74	The Green Knight
M20	1995	76	Jackson's Dilemma

AGATHA CHRISTIE

TECHNOLOGY[1]	I.D.	YEAR OF PUBLICATION	APPROX. AGE AT COMPOSITION	NOVEL
Typewriter	C1	1920	28	The Mysterious Affair at Styles
	C2	1922	32	The Secret Adversary
	C3	1926	34	The Murder of Roger Ackroyd
	C4	1934	43	Murder on the Orient Express
	C5	1937	47	Appointment with Death
	C6	1975	50	Curtain[2]
	C7	1944	51	Towards Zero
	C8	1950	59	A Murder is Announced
Dictaphone	C9	1954	63	Destination Unknown
	C10	1958	67	Ordeal by Innocence
	C11	1963	72	The Clocks
	C12	1967	76	Endless Night
	C13	1970	79	Passenger to Frankfurt
	C14	1971	80	Nemesis
	C15	1972	81	Elephants Can Remember
Dictaphone + Editing	C16	1973	82	Postern of Fate

[1] Lancashire (Forthcoming 2010).

[2] Written between 1940 and 1941.

P. D. James

I.D.	Year of Publication	Approx. Age at Composition	Novel
J1	1962	42	Cover Her Face
J2	1963	43	A Mind to Murder
J3	1967	47	Unnatural Causes
J4	1971	51	Shroud for a Nightingale
J5	1972	52	An Unsuitable Job for a Woman
J6	1975	55	The Black Tower
J7	1977	57	Death of an Expert Witness
J8	1980	60	Innocent Blood
J9	1986	66	Taste for Death
J10	1992	72	The Children of Men
J11	1997	77	A Certain Justice
J12	2001	81	Death in Holy Orders
J13	2003	83	The Murder Room
J14	2005	85	The Lighthouse
J15	2008	88	The Private Patient

B Methods: Further Discussion

B.1 Dialogue Detection

The texts used in our analysis are novels, which inevitably contain a mixture of narrative and dialogue. Natural dialogue is often characterized by a significant number of fragments and elliptical constructions, since speakers may interrupt each other or reply to questions with brief responses. To simulate natural speech, fiction writers sometimes alter their writing styles and develop ways of speaking unique to each character. Dialogues, therefore, are not good indicators of the syntactic level nor (though arguably to a lesser extent) the lexical level of a writer.

Ideally, dialogues should be analyzed separately from narratives, if at all. A naïve algorithm to filter out dialogues may proceed as follows: the first quotation mark encountered in a text (double or single, depending on the typography) signals the beginning of direct speech, the second quotation mark signals the end, and the process continues in this fashion for subsequent odd-numbered and even-numbered quotation marks; if single quotation marks are used, they must be distinguished from apostrophes. In practice, this process is not as straightforward. First, the separation is not always clear-cut: dialogue and narrative can be intertwined. For instance,

> Before this final stage of his illness had fallen upon him, Simon Maxie had whispered to her, " You won't let them take me away, Eleanor? " and she had replied, " Of course I won't. " [J1]

In the above example, we use straight quotation marks surrounded by spaces for the following reasons. The OCR software we used was configured to reproduce the novels in plain-text format for automated analysis; thus the curved quotation marks in the original texts were converted to straight quotation marks. Correct spacing before and after each quotation mark cannot be relied upon: because of the uneven spacing due to text justification, extra whitespaces may be added.[1] These lead to the second problem: while the common typography for nested quotations is to alternate between double and single quotation marks, this format cannot be guaranteed. When the same

[1] Neither is correct spacing guaranteed for other punctuation marks and whole words; however, this does not pose a problem for our algorithm.

type of quotation marks is used, some portion(s) of the quotation will be erroneously identified as narrative, such as the underlined word in the following example:

> " Why doesn't she say " Stephen " ? " thought Mrs Maxie irrelevantly. [J1]

Had correct spacing been ensured, this problem could have been rectified: no ambiguity will arise if an opening quotation mark is attached to the first word of the quotation, and a closing quotation mark to the last.

Furthermore, the text portion in between quotation marks is not necessarily a part of a dialogue. Aside from signaling verbatim speech, quotation marks can be used to mark irony, unusual word usage, titles of creative works, or use-mention distinction. The following passage demonstrates some examples of such usage:

> I wonder if I shall ever write my Charles Arrowby Four Minute Cookbook? The ' four minutes ' of course refer to the active time of preparation, and do not include unsupervised cooking time. I have looked at several so-called ' short order ' cookery books, but these works tend to deceive, their ' fifteen minutes ' really in practice means thirty, and they contain instructions such as ' make a light batter ' . [M15]

The final complication is dialogues spanning several paragraphs, for which the convention is to omit the closing quotation marks of all paragraphs but the very last one, for instance:

> " <*first paragraph*>
>
> " <*second paragraph*>
>
> " <*last paragraph*> "

The naïve separation technique, which marks the second paragraph as narrative, fails to handle this case. If correct punctuation and line breaks were preserved, the algorithm could be modified to treat any paragraph that begins with a quotation mark but ends without one as a part of an ongoing multi-paragraph quotation. However, our data might not meet these conditions, considering the possible errors in the digitizing process: dust on the pages may be mistaken for quotation marks; fuzzy printing can cause quotation marks to be ignored or incorrectly identified as numbers in superscript; and an extra line break is added at the end of each physical page in the printed materials.

Despite our data correction process, absolute accuracy cannot be guaranteed, and one missing or extra quotation mark will invert the narrative/dialogue detection results from that point onwards. Because of this severe impact on accuracy, we analyzed the texts without distinction between dialogue and narrative. Consequently, the overall syntactic scores of the authors may be lower, compared to a narrative-only analysis, given the higher proportions of fragments, ellipses and simple sentences characteristic of dialogue. A similar discrepancy may exist in the lexical analysis, since the authors may have altered their word usage to reflect the origins, education levels, social classes, or cultural backgrounds of their characters. The extent of the impact of including dialogue in the analysis cannot be determined until the narrative-only analysis is carried out.

B.2 WordNet Depth as Specificity Rank

As documented in section 4.1, we relied on WordNet depths to approximate the average noun specificity rank of each text. The basis for this approximation is the hypernym–hyponym-based organization of WordNet's noun hierarchy. The same does not apply to other content word classes. Adjectives and adverbs are placed along bipolar scales between pairs of extremes, rather than in hierarchies. Although verbs are structured based on their hypernym–troponym relations, unlike nouns, they are organized into several wide and shallow hierarchies with different roots. The depth of a verb synset in one hierarchy may not be comparable to the depth of another synset in a different hierarchy. For instance, the first synset of the verb *express* (denoted $express_1$), which means to show, to "give expression to," has a WordNet depth of 7 from the root act_1:

$$\text{act}_1 > \text{interact}_1 > \text{communicate}_2 > \text{inform}_1 > \text{tell}_2 > \text{impart}_1 > \text{convey}_1 > \mathbf{express_1}$$

while its second synset, $express_2$, (to "articulate; either verbally or with a cry, shout, or noise") has depth 0, since it is the root of a separate hierarchy. A direct troponym of $express_2$, at depth 1, is say_1 (to "utter aloud"). Considered together, the depths of $express_1$, $express_2$, say_1 and $tell_2$ (which are 7, 0, 1 and 4, respectively) do not reflect the relative degrees of specificity among these verbs. WordNet verb hierarchies are therefore unsuitable for our purpose.

The structure for nouns, with one single hierarchy rooted at *entity*$_1$, provides greater consistency than a multiple-hierarchy structure; however, it is not without problems. As discussed in section 4.3, there may be more than one path from the root *entity*$_1$ to any given synset, for instance:

entity$_1$ > physical_entity$_1$ > object$_1$ > whole$_2$ > living_thing$_1$ > organism$_1$ > animal$_1$
> domestic_animal$_1$ > **dog$_1$**

entity$_1$ > physical_entity$_1$ > object$_1$ > whole$_2$ > living_thing$_1$ > organism$_1$ > animal$_1$
> chordate$_1$ > vertebrate$_1$ > mammal$_1$ > placental$_1$ > carnivore$_1$ > canine$_2$ > **dog$_1$**

The difference between these two paths is that, at the node *animal*$_1$, there are two sub-branches through either *domestic_animal*$_1$ or *chordate*$_1$, one containing more fine-grained distinctions than the other. As a result, *dog*$_1$ has a maximum depth of 13, while its minimum depth is 8. The synset *cat*$_1$, which is of the same semantic category and at the same specificity level as *dog*$_1$, has exactly one path of depth 13 from the root *entity*$_1$:

entity$_1$ > physical_entity$_1$ > object$_1$ > whole$_2$ > living_thing$_1$ > organism$_1$ > animal$_1$
> chordate$_1$ > vertebrate$_1$ > mammal$_1$ > placental$_1$ > carnivore$_1$ > feline$_1$ > **cat$_1$**

Because *cat*$_1$ lacks a "shortcut" through *domestic_animal*$_1$, its minimum depth is much greater than that of its semantic neighbour *dog*$_1$. One could then argue that WordNet's maximum depth is a better approximation of specificity. A problem associated with this approach, however, is that branches in WordNet are not equally fine-grained. While *dog*$_1$ and *cat*$_1$ have maximum depths of 13, nouns specifying humans and their occupations have much smaller maximum depths. Among the nouns taken from the first few paragraphs of P. D. James's *Cover Her Face* (see Table 21), *vicar*$_2$, *expert*$_1$ and *man*$_1$ have maximum depths of 10, 7 and 8, respectively. However, *expert*$_1$ is arguably more specific than *man*$_1$, and all of these should be at least as specific as *cat*$_1$ and *dog*$_1$.

As shown in Table 21, WordNet depths reflect well the relative specificity among nouns of similar or related categories, such as: *man*$_1$–*woman*$_1$–*girl*$_1$, *table*$_2$–*cupboard*$_1$, *meal*$_1$–*dinner*$_{1,2}$. However, problems arise for words belonging to distinct categories; some arguably problematic cases are: *coffee*$_1$ being more specific than *brain*$_3$, *baby*$_1$ having the same rank as *vicar*$_2$, and *table*$_2$ being twice as specific as *wisdom*$_2$ in terms of both maximum and minimum depths. Considering the inconsistencies observed in our WordNet-based noun specificity results (presented in section 5), we conclude that WordNet depth is unsuitable as a source for word specificity approximation.

Table 21: Examples of WordNet depths for different noun synsets

Noun token	Synset#	Depth		Definition
		Max.	Min.	
ABSTRACT ENTITY				
accuracy	1	4	4	the quality of being near to the true value
aura	3	4	4	a distinctive but intangible quality surrounding a person or thing
erudition	1	6	6	profound scholarly knowledge
wisdom	2	4	4	the quality of being prudent and sensible
PHYSICAL ENTITY				
baby	1	10	7	a very young child (birth to 1 year) who has not yet begun to walk or talk
brain	3	4	4	that which is responsible for one's thoughts and feelings; the seat of the faculty of reason
coffee	1	8	6	a beverage consisting of an infusion of ground coffee beans
cupboard	1	8	8	a small room (or recess) or cabinet used for storage space
digestion	1	5	5	the process of decomposing organic matter (as in sewage) by bacteria or by chemical action or heat
dinner	1	7	7	the main meal of the day served in the evening or at midday
dinner	2	7	7	a party of people assembled to have dinner together
evening	1	6	6	the latter part of the day (the period of decreasing daylight from late afternoon until nightfall)
expert	1	7	4	a person with special knowledge or ability who performs skillfully
girl	1	9	6	a young woman
man	1	8	5	an adult person who is male (as opposed to a woman)
meal	1	6	6	the food served and eaten at one time
mother	1	12	9	a woman who has given birth to a child (also used as a term of address to your mother)
table	2	8	8	a piece of furniture having a smooth flat top that is usually supported by one or more vertical legs
vicar	2	10	7	a clergyman in charge of a chapel
woman	1	8	5	an adult female person (as opposed to a man)

B.3 Pattern Sets

Passive:

A passive sentence is one that matches any of the following patterns, which were designed for use in the branch-matching mode.

Table 22: Patterns for explicit forms of passive voice

Type	Pattern specification

be-passive	`([VP SQ] ([AUX AUXG] [be am is are being was were been])`
	`(VP ([VBN VBD])))`

get-passive	`(VP ([VB VBP VBZ VBG VBD VBN] [get gets getting got gotten])`
	`(VP ([VBN VBD])))`

by-phrase	`(VP (VBN) (PP (IN by) (NP)))`

D-Level:

Table 23 displays the patterns developed for the D-Level scale, all of which are for the branch-matching mode unless otherwise indicated. Complicated patterns are illustrated with example matching sentences; these examples were given by Covington et al. (2006) in their revised D-Level scale, except for some additional examples for Levels 4 and 6.

Frequently repeated portions of the patterns are defined separately and are substituted into the main patterns that use them. The names of these substituted terms are headed by an ampersand to distinguish them from normal pattern tags and values. For example, the term `&noun` stands for the string `NP NN NNS NNP NNPS PRP`, and is substituted into patterns that involve `&noun`, such as `(NP ([&noun]))`, which then becomes `(NP ([NP NN NNS NNP NNPS PRP]))`. Table 24 lists all such terms and their corresponding substitution strings. Table 25 presents a special case: the substitution strings are lists of verbs in base forms; conjugated forms of these verbs are generated automatically and then included in the substitutions. These verb lists, grouped by the type of complement that they take, are drawn from Huddleston and Pullum (2002).

Table 23: Patterns for the revised D-Level scale

LEVEL 0	Simple sentence (root-matching mode)	`([S SINV] (VP))`
	Sentences with auxiliaries and semi-auxiliaries	`([MD AUX AUXG])`
	Question and fragment (root-matching mode)	`([SQ SBARQ NP VP ADJP ADVP PP FRAG])`

LEVEL 1	Infinitive or -*ing* complement with same subject as main clause
	e.g., *Try to brush her hair.*
	Try brushing her hair.
	I felt like turning it.
		`(VP =([&verb] [&vp-to-exceptions &raising-verb]) (S +(VP (TO) (VP ([VB VBP])))))`
		`(VP ([&verb]) ~(ADVP) (S +(VP (VBG))))`
		`(VP ([&verb]) ~(ADVP) (S +(VP ^(VP (VBG)))))`
		`(VP ([&verb]) (PP (IN like) (S +(VP (VBG)))))`
		`(VP ([&verb]) (PP (IN like) (S +(VP ^(VP (VBG))))))`

LEVEL 2	Conjoined noun phrases in subject position	`([&cls] (NP ([&noun]) (CC) ([&noun])))`
	Sentences conjoined with a coordinating conjunction	`([&cls] ([&cls]) (CC) ([&cls]))`

	Conjoined verbal, adjectival, or adverbial constructions	`(VP ([&verb]) (CC) ([&verb]))`
		`(ADJP ([&adj]) (CC) ([&adj]))`
		`(ADVP ([&adv]) (CC) ([&adv]))`

LEVEL 3	Relative (or appositional) clause modifying object of main verb
	e.g., *The man scolded the boy who stole the bicycle.*
		`(VP *(NP ([&noun]) (SBAR ~(IN that) (S))))`
		`(VP *(NP ([&noun]) (SBAR (WHNP (IN that)) (S))))`
		`(VP (NP) ~(,) (SBAR ~(IN that) (S)))`
		`(VP *(_ (NP) ~(,) (SBAR ~(IN that) (S))))`
		`(VP *(NP ([&noun]) (SBAR ([WHNP WHPP WHADVP]) (S))))`
		`(VP ([&verb]) +(SBAR (WHNP) (S)))`
		`(VP ([&verb]) *(PP (IN) +(SBAR ([WHNP WHPP WHADVP]) (S))))`

	Nominalization in object position
	e.g., *Why can't you understand his rejection of the offer?*
		`(VP ([&nominalization-patterns]))`
		`(VP ([NP PP] ([&nominalization-patterns])))`
		`(VP ^(NP (POS)) (S (VP (VBG) (NP))))`

	Finite clause as object of main verb
	e.g., *John knew that Mary was angry.*
	Remember where it is?
		`(VP ([&verb]) ~([ADVP PP PRT ,]) (SBAR *(S ([&finite-vp]))))`
		`(VP ([&verb]) +([&cls] (SBAR *(S ([&finite-vp])))))`
		`(SINV ([PP ADJP]) (VP) (NP))`

	Raising e.g., *John seems to Mary to be happy.*	`(S (NP) ^(VP ([&finite-verb] [&raising-verb]) ~(PP) (S (VP (TO) (VP (VB))))))`

Level 3	Subject extraposition
	e.g., *It was surprising for John to have left Mary*.
		`([&cls] (NP (PRP it)) ^(VP ([&verb])`
		`([NP ADJP PP]) (SBAR (S (VP)))))`
		`([&cls] (NP (PRP it)) (VP *(VP ([&verb])`
		`([NP ADJP PP]) (SBAR (S (VP))))))`

Level 4	Comparative with object of comparison
	e.g., *John is older than Mary*.
		`([ADJP NP] ^([ADJP NP] ([JJR RBR]))`
		`([(PP (IN than) ([&noun])) (SBAR (IN than) (S))]))`

	Non-finite complement with its own understood subject, with some forms of relative clause	

	– Constructions with embedded non-finite verb phrase*
	e.g., *I expect him to go.*
	I want it done today.
	I saw him walking the dog.
		`(VP ([&verb]) ~([ADVP PRT])`
		`(S (NP ([&noun])) ([&nonfin-vp])))`
		`(VP ([&verb]) ~([ADVP PRT])`
		`(NP ([&noun]) =(POS)) (S ([&nonfin-vp])))`
		`(SQ (VP ([&verb])) (NP) ([&nonfin-vp]))`
		`(SQ (VP ([&verb]) (NP) ([&nonfin-vp])))`

	–Constructions with predicative complements (PC)	

	— Main verb taking adjective as PC
	e.g., *This problem drives me mad.*
	The cops held my neighbour's son responsible.
	It was my sister he considered pretty.
		`(VP ^([&verb] [&vp-pc &vp-adj &vp-opt-pc &vp-opt-as-pc])`
		`+(S +(NP) +(ADJP)))`
		`(VP ^([&verb] [&vp-pc &vp-adj &vp-opt-pc &vp-opt-as-pc])`
		`+(NP ~([&mod-noun]) ([&noun]) ~([POS &noun]) ([&adj])))`
		`([SBAR SBARQ] ~([WHNP (IN that)])`
		`([S SQ] (NP)`
		`^(VP ^([&verb] [&vp-pc &vp-adj &vp-opt-pc &vp-opt-as-pc])`
		`+([ADJP (S +(ADJP))])))))`

	— Main verb taking noun phrase as PC
	e.g., *I consider John a friend.*
	I'm inviting John, whom I consider a friend.
		`(VP ^([&verb] [&vp-pc &vp-np &vp-opt-pc &vp-opt-as-pc])`
		`+(NP) +(NP))`
		`(VP ^([&verb] [&vp-pc &vp-np &vp-opt-pc &vp-opt-as-pc])`
		`+(S +(NP) +(NP) =([&verb])))`
		`([SBAR SBARQ] ~([WHNP (IN that)])`
		`([S SQ] (NP)`
		`^(VP ^([&verb] [&vp-pc &vp-np &vp-opt-pc &vp-opt-as-pc])`
		`+(NP ([&noun])))))`

	— Main verb taking infinitival complements, pattern for relative clause. (Original form handled in [*])
	e.g., *Meet John, the man they elected to be president.*
		`([SBAR SBARQ] ~([WHNP (IN that)])`
		`([S SQ] (NP)`
		`^(VP ^([&verb] [&vp-inf])`
		`(S (VP (TO) (VP (VB)`
		`+([ADJP (S +(ADJP)) (NP)]))))))))`

LEVEL 4	— Main verb requiring *as*, taking noun phrase as PC
	e.g., *They elected John as president.*
	Meet John, the man they elected as president.
		`(VP ^([&verb] [&vp-np]) +(NP ([&noun])) +(PP (IN as) (NP ([&noun]))))`
		`([SBAR SBARQ] ~([WHNP (IN that)]) ([S SQ] (NP) ^(VP ^([&verb] [&vp-np]) +(PP (IN as) (NP ([&noun]))))))`

	— Main verb requiring *as*, taking either noun phrase or adjective as PC
	e.g., *I regard her as my best friend.*
	I regard her as indispensable.
	They fired Sue, whom I regard as indispensable.
		`(VP ^([&verb] [&vp-as-pc &vp-opt-as-pc]) +(NP ([&noun])) +(PP (IN as) ([ADJP NP])))`
		`([SBAR SBARQ] ~([WHNP (IN that)]) ([S SQ] (NP) ^(VP ^([&verb] [&vp-as-pc &vp-opt-as-pc]) +(PP (IN as) ([ADJP NP])))))`
		`([SBAR SBARQ] ~([WHNP (IN that)]) ([S SQ] (NP) ^(VP ^([&verb] [&vp-as-pc &vp-opt-as-pc]) +(NP) +(PP (IN as)))))`

LEVEL 5	Sentences joined by a subordinating conjunction
	e.g., *They will play today if it does not rain.*
		`([&cls] *([&subor-conj-phrase]))`
		`(VP *([&subor-conj-phrase]))`

	Nonfinite clauses in adjunct (not complement) positions
	e.g., *Cookie Monster touches Grover after jumping over the fence.*
	Having tried both, I prefer the second one.
		`(PP (IN [after before despite once when whether while]) (S (VP ([&nonfin-verb]))))`
		`([&cls] ([&nonfin-cls]) (, ,))`
		`(_ (, ,) ([&nonfin-cls]))`
		`([&cls] (S ([&nonfin-vp]) (VP)))`

LEVEL 6	Relative clause modifying subject of main verb
	e.g., *The man who cleans the rooms left early.*
		`([&cls] (NP ([&noun]) *(SBAR ~(IN that) (S))) (VP))`
		`([&cls] (NP) ~(,) (SBAR ~(IN that) (S)) (VP))`
		`([&cls] (NP ([&noun]) *(SBAR ([WHNP WHPP WHADVP]) (S))) (VP))`

	Appositional clause modifying subject of main verb
	e.g., *John, the man who cleans the rooms, left early.*
	John—the man who cleans the rooms—left early.
		`([&cls] (NP +(NP) +(,) +(NP) +(,)) (VP))`
		`([&cls] (NP +(NP) +(: -) +(NP) +(: -)) (VP))`
		`([&cls] (NP +(NP) +(: -) +([&nonfin-vp]) +(: -)) (VP))`
		`([&cls] (NP *(PRN ([&cls]))))`

	Embedded clause serving as subject of main verb	`([&cls] ([&cls]) +(VP))`

	Nominalization serving as subject of main verb
	e.g., *John's refusal of the drink angered Mary.*
		`([&cls] ([&nominalization-patterns]) +(VP))`
		`([&cls] (NP ([&nominalization-patterns])) +(VP))`

LEVEL 7	More than one level of embedding in a single sentence e.g., *John decided to leave Mary when he heard that she was seeing Mark.*	(Any combination of the above patterns from at least 2 different levels.)

Table 24: Substitutions for D-Level patterns

Term	Substitution string
&cls	S SBAR SBARQ SINV SQ
&noun	NP NN NNS NNP NNPS PRP
&nom	NN NNS NNP NNPS PRP VBG
&verb	VP VB VBD VBG VBN VBP VBZ MD AUX AUXG
&nonfin-verb	VB VBG VBN AUXG
&finite-verb	VBD VBP VBZ MD AUX
&adj	ADJP JJ JJR JJS
&adv	ADVP RB RBR RBS WRB
&mod-noun	PDT DT PRP$ ADJP JJ JJR JJS
&mod-verb	ADVP PP
&vp-to-exceptions	be am is are was were been bound have going got need supposed used
&fin-vp	(VP ~([&mod-verb]) ([VBD VBP VBZ MD AUX]))
&finite-vp	&fin-vp (VP &fin-vp (CC) &fin-vp)
&nonfin-vp	(VP ([VB VBG VBN AUXG VBP])) (VP (TO to) ^(VP ([VB VBP])))
&nonfin-cls	(S ([&nonfin-vp])) (S (S ([&nonfin-vp])) (CC)) (S (CC) (S ([&nonfin-vp]))) (_ ([&noun]) +([ADJP PP NP]) -([&verb])) (S (_ ([&noun]) +([ADJP PP NP]) -([&verb])))
&subor-conj-phrase	(SBAR (IN [after although as because before for if lest once since though till unless until whereas whether while]) (S ([&finite-vp]))) (SBAR (_ (WRB [when whenever wherever])) (S)) (SBAR (IN so) +(IN that) (S)) (SBAR (IN in) +(IN that) (S)) (_ *(_ [so in now]) +(SBAR +(IN that) (S))) (SBAR (IN in) (NN case) (S)) (SBAR (IN in) (NN order) (S)) (PP (IN by) (NP (NP (DT the) (NN time)) (SBAR))) (_ (NP (DT every) (NN time)) (SBAR)) (_ (IN in) (NP (DT the) (NN event) (SBAR))) (_ (DT no) (NN matter) (SBAR ([WHADVP WHNP]) (S))) (_ (_ (provided providing supposing)) ([S SBAR]))
&nominalization-patterns	(NP (NP ([DT PRP$ (NP (POS))]) *([&nom] is_nom)) (PP (IN) ([NP (S (VP (VBG)))]))) ([NP S] ([DT PRP$ (NP) (NP (POS))]) +(VP ([VBG AUXG])))

Table 25: Auto-generated substitutions for D-Level patterns

Term	Base string
&raising-verb	seem appear begin continue
&vp-pc	believe certify consider declare deem feel find judge like prefer presume profess pronounce prove reckon report rule think want account brand call designate esteem hold imagine keep label leave rate term get make render
&vp-inf	believe certify consider declare deem feel find hold judge like prefer presume profess pronounce prove reckon report rule think want appoint designate elect proclaim
&vp-adj	have hold wish drive put send set turn
&vp-np	appoint baptise baptize christen create crown designate elect name proclaim vote
&vp-refx	acknowledge confess suppose
&vp-opt-pc	boil bore brush drain fill frighten jerk plane shoot wash knock paint rub push wipe
&vp-as-pc	accept acknowledge adopt bill brand cast categorise categorize characterise characterize choose class classify condemn confirm construe count define denounce depict describe diagnose disguise dismiss enlist establish give hail have identify instal install intend interpret know mean perceive portray present recognise recognize regard represent scorn see suggest take treat use view
&vp-opt-as-pc	consider imagine nominate ordain rate report

References

Elizabeth Bates, Christine Harris, Virginia Marchman, Beverly Wulfeck, and Mark Kritchevsky. Production of complex syntax in normal ageing and Alzheimer's disease. *Language and Cognitive Processes*, 10:487–539, 1995.

Helen Bird, Matthew A. Lambon Ralph, Karalyn Patterson, and John R. Hodges. The rise and fall of frequency and imageability: Noun and verb production in semantic dementia. *Brain and Language*, 73:17–49, 2000.

Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly, 1st edition, 2009. http://www.nltk.org/book.

Dan G. Blazer and David C. Steffens. *The American Psychiatric Publishing Textbook of Geriatric Psychiatry*. American Psychiatric Publishing, Inc., 4th edition, 2009.

Eugene Charniak. Charniak parser, 2006. http://www.cs.brown.edu/~ec/.

Hintat Cheung and Susan Kemper. Competing complexity metrics and adults' production of complex sentences. *Applied Psycholinguistics*, 13:53–76, 1992.

Agatha Christie. *Agatha Christie: An Autobiography*. Dodd, Mead & Company, New York, 1977.

Michael A. Covington, Congzhou He, Cati Brown, Lorina Naci, and John Brown. How complex is that sentence? A proposed revision of the Rosenberg and Abbeduto D-Level Scale. Research Report 2006-01, CASPR, 2006. http://www.ai.uga.edu/caspr/2006-01-Covington.pdf.

Peter Garrard, Lisa M. Maloney, John R. Hodges, and Karalyn Patterson. The effects of very early Alzheimer's disease on the characteristics of writing by a renowned author. *Brain*, 128(2): 250–260, 2005.

William C. Groves, Jason Brandt, Martin Steinberg, Andrew Warren, Adam Rosenblatt, Alva Baker, and Constantine G. Lyketsos. Vascular dementia and Alzheimer's disease: Is there a difference? A comparison of symptoms by disease duration. *Journal of Neuropsychiatry and Clinical Neurosciences*, 12:305–315, 2000.

Hedvig Holm, Maria Mignéus, and Elisabeth Ahlsén. Linguistic symptoms in dementia of Alzheimer type and their relation to linguistic symptoms of aphasia. *Logopedics Phoniatrics Vocology*, 19(3):99–106, 1994.

Rodney Huddleston and Geoffrey K. Pullum. *The Cambridge Grammar of the English Language*. Cambridge University Press, 1st edition, 2002.

Susan Kemper and Aaron Sumner. The structure of verbal abilities in young and older adults. *Psychology and Aging*, 16:312–322, 2001.

Susan Kemper, Donna Kynette, Shannon Rash, Kevin O'Brien, and Richard Sprott. Life-span changes to adults' language: Effects of memory and genre. *Applied Psycholinguistics*, 10(1): 49–66, 1989.

Susan Kemper, Marilyn Thompson, and Janet Marquis. Longitudinal change in language production: Effects of aging and dementia on grammatical complexity and propositional content. *Psychology and Aging*, 16:600–614, 2001.

Ian Lancashire. *Forgetful Muses: Reading the Author in the Text*. University of Toronto Press, Toronto, Forthcoming 2010.

Ian Lancashire and Graeme Hirst. Vocabulary changes in Agatha Christie's mysteries as an indication of dementia: A case study. 2009. 19th Annual Rotman Research Institute Conference, Cognitive Aging: Research and Practice, 8–10 March 2009, Toronto. http://ftp.cs.toronto.edu/pub/gh/Lancashire+Hirst-extabs-2009.pdf.

Jane Maxim and Karen Bryan. *Language of the Elderly: A Clinical Perspective*. Whurr, London, 1994.

Janet Morgan. *Agatha Christie: A Biography*. Collins, London, 1984.

Marjorie Nicholas, Loraine K. Obler, Martin L. Albert, and Nancy Helm-Estabrooks. Empty speech in Alzheimer's disease and fluent aphasia. *Journal of Speech and Hearing Research*, 28: 405–410, 1985.

Ted Pedersen. WordNet::SenseRelate, 2009. http://www.d.umn.edu/~tpederse/senserelate.html.

Holly B. Posner, Ming-Xin Tang, Jose Luchsinger, Rafael Lantigua, Yaakov Stern, and Richard Mayeux. The relationship of hypertension in the elderly to AD, vascular dementia, and cognitive function. *Neurology*, 58:1175–1181, 2002.

Sheldon Rosenberg and Leonard Abbeduto. Indicators of linguistic competence in the peer group conversational behavior of mildly retarded adults. *Applied Psycholinguistics*, 8:19–32, 1987.

Lorna Sage. *The Cambridge Guide to Women's Writing in English*. Cambridge University Press, Cambridge, 1999.

Shanne R. Smith, Helen J. Chenery, and Bruce E. Murdoch. Semantic abilities in dementia of the Alzheimer type: II. Grammatical semantics. *Brain and Language*, 36:533–542, 1989.

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

Laura Thompson. *Agatha Christie: An English Mystery*. Headline Review, London, 2007.

Kristine Williams, Frederick Holmes, Susan Kemper, and Janet Marquis. Written language clues to cognitive changes of aging: An analysis of the letters of King James VI/I. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 58B(1):42–44, 2003.

Victor H. Yngve. A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society*, 104(5):444–466, 1960.