Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three British novelists

Xuan Le
Department of Computer Science, University of Toronto, Toronto, Ontario, Canada M5S 3G4

Ian Lancashire
Department of English, University of Toronto, Toronto, Ontario, Canada M5R 2M8

Graeme Hirst
Department of Computer Science, University of Toronto, Toronto, Ontario, Canada M5S 3G4

Regina Jokel
Department of Speech–Language Pathology, University of Toronto, Toronto, Ontario, Canada M5G 1V7 and Kunin-Lunenfeld Applied Research Unit, Baycrest Hospital, 3560 Bathurst Street, North York, Ontario, Canada M6A 2E1

Abstract

We present a large-scale longitudinal study of lexical and syntactic changes in language in Alzheimer's disease using complete, fully parsed texts and a large number of measures, using as our subjects the British novelists Iris Murdoch (who died with Alzheimer's), Agatha Christie (who was suspected of it), and P.D. James (who has aged healthily). We avoid the limitations and deficiencies of Garrard *et al.*'s [(2005), The effects of very early Alzheimer's disease on the characteristics of writing by a renowned author. *Brain*, 128 (2): 250–60] earlier study of Iris Murdoch. Our results support the hypothesis that signs of dementia can be found in diachronic analyses of patients' writings, and in addition lead to new understanding of the work of the individual authors whom we studied. In particular, we show that it is probable that Agatha Christie indeed suffered from the onset of Alzheimer's while writing her last novels, and that Iris Murdoch exhibited a 'trough' of relatively impoverished vocabulary and syntax in her writing in her late 40s and 50s that presaged her later dementia.

Correspondence:
Graeme Hirst
Department of Computer Science, University of Toronto, Toronto, Ontario, Canada M5S 3G4.
E-mail: gh@cs.toronto.edu

Literary and Linguistic Computing, Vol. 26, No. 4, 2011.

doi:10.1093/llc/fqr013 Advance Access published on 24 May 2011

1 Introduction

The research we describe in this article has two complementary goals. First, we want to determine whether linguistic markers of Alzheimer's disease can be detected in a diachronic analysis of aspects of an individual's writing. Our subjects for this are novelists for whom we have a large, lifelong body of work. Second, with regard to the individual authors who serve as subjects, we aim to conduct much larger and more extensive studies than those of prior research, correcting the deficiencies of Garrard *et al.*'s (2005) study of Iris Murdoch, by using a large corpus and state-of-the-art analysis methods.

1.1 Motivation, background, and approach

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, involving several diagnostic procedures, may be highly stressful for the individuals undergoing diagnosis—yet a definitive diagnosis can be made only post mortem. But while there is no proven cure for many types of dementia, a correct, timely diagnosis is of great importance; in the future, a sufficiently early diagnosis of AD may even make prevention possible, according to Blazer and Steffens (2009): '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).

Recent studies further suggest that early diagnosis can also be achieved through linguistic analysis.[1] The fact that the disease negatively affects the linguistic abilities of patients in both speech and writing presents the possibility of developing non-intrusive evaluation techniques that require minimal involvement from the patients, by looking for diachronic changes in their writing. If a patient's lifelong corpus of writing is available in an online format, as is likely for future generations, these techniques could be used in conjunction with clinical assessments or on their own as an early detection tool.

To test the hypothesis that the disease is detectable by a diachronic analysis of a patient's writing, we need writing samples from both healthy elderly adults and diagnosed dementia patients. 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 present-day 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 the lead of Garrard *et al.* (2005) (whose work and its limitations we discuss in Section 1.3 below), we chose the writings of Iris Murdoch, who died with AD, as the linguistic model of dementia patients and contrasted it with the linguistic model of healthily aging adults represented by the writings of crime fiction author P.D. James, who, in 2008 at 88 years, her powers undiminished and her health not in doubt, published her latest novel, *The Private Patient*. Expanding on work by Lancashire (2010; Lancashire and Hirst 2009) (see Section 1.3), we also analyzed the novels of Agatha Christie, which present an interesting case study of possible undiagnosed dementia.

1.2 Language in aging and dementia

Alzheimer's 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 (Blazer and Steffens, 2009). Healthily aging adults may also experience a decline in their cognitive abilities, albeit one that is significantly less severe, and changes in their language (Maxim and Bryan, 1994; Burke and Shafto, 2008).

Here, we summarize lexical and syntactic changes in healthy aging and in dementia; a more-detailed discussion is given by Le (2010). Kemper *et al.* (2001) and Burke and Shafto (2008) report that in healthy aging, vocabulary increases through the middle adult years, but then may start to decline. In dementia, vocabulary declines much more

rapidly, especially the use of low-frequency and more-specific words (Maxim and Bryan, 1994; Bird *et al.*, 2000; Burke and Shafto, 2008), a consequence of which is that the patient's noun-to-verb ratio changes as more low-image verbs are used (Bird *et al.*, 2000). Moreover, lexical repetitions increase (Nicholas *et al.*, 1985; Smith *et al.*, 1989; Holm *et al.*, 1994; Cook *et al.*, 2009); 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); fillers (*um, ah*) and dysfluencies increase (Burke and Shafto, 2008).

The syntactic complexity of language, defined by measures such as clauses per utterance, declines with age in both spoken and written language (Burke and Shafto, 2008). Maxim and Bryan (1994) report that left-branching clauses in English are more difficult for elderly adults to process than for a younger control group. Kemper *et al.* (2001), in a longitudinal study following linguistic changes in healthy elders and dementia patients, found decline in grammatical complexity to be far more rapid in the latter. Bates *et al.* (1995) found that use of the passive voice, in particular, was affected, with healthy elders producing fewer passives than a younger control group, and Alzheimer's patients far fewer again. Moreover, the AD group used more agentless passives (e.g., *John was fired* or *John got fired*) than either of the control groups, and also relied heavily on the *get* form of the passive.

In summary, while heterogeneity is expected in the linguistic changes among individuals in both normal aging and dementia, and 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.

1.3 Related work

In recent years, several longitudinal studies have been conducted with focus on individual writers, in order to examine their patterns of linguistic changes over time. For example, 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 vascular dementia. The study did not produce a conclusive diagnosis.

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 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 her declining cognitive health.[2] Along with this novel, Garrard *et al.* 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*. However, the linguistic analysis conducted suffered from serious problems in methodology (see Le, 2010 for a more-detailed critique).

The first problem is the data used for analysis, which consisted of the complete texts of *Under the Net* and *Jackson's Dilemma,* but, for unexplained reasons, only about 20% of *The Sea, the Sea* (the first 100 pages). Length-based measures were extrapolated from this sample on the questionable assumption that the remaining 80% of the novel was identical in structure to the first 20% (indeed, we found, in our experiments below, that the projected word count was off by nearly 7000 tokens).

Then, despite having 2.2 full novels available, much of the analysis was just carried out manually on tiny samples from the texts. Average word length, average word frequency, and grammatical class proportions were computed from five 100-word samples from each novel. The results of these measures are therefore not reliable. Moreover, the grammatical class of each word was determined

not by conventional context-dependent part-of-speech tagging, but by simply selecting ''the more typical reading'' out of four categories—noun, verb, descriptor, and function word—regardless of context (p. 253). The accuracy of this approximation is uncertain, since relatively few English words belong to only one word class. For the remaining measures performed by hand, Garrard *et al.* 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).

Garrard *et al.*'s remaining measures (on the complete 2.2 texts) 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 of *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*. Their final measure, termed auto-collocations, computes the proportion of times the ten most common words were repeated within a space of four subsequent words. Since the most common repetitions are inevitably function words, this measure was classified as a syntactic analysis technique; however, little conclusion can be drawn from it with regards to syntactic complexity. 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 the small samples randomly drawn from the three novels.[3]

In work that led to the project to be described in this article (and whose results are incorporated into Section 3 below), Lancashire (2010; Lancashire and Hirst 2009) analyzed 15 complete novels by Agatha Christie from 1922 to 1973 in late 2007 and detected, in her last novels (especially *Elephants Can Remember* and *Postern of Fate*), both written in her early 80s, a sizable loss of vocabulary as well as large increases in phrasal repetitions and indefinite nouns. Both novels work out the clues and plot of their murders with difficulty; and the female leading character of *Elephants* writes detective fiction and complains of a deteriorating memory. Biographies suggest that in Christie's 80s, her daughter forbade her publisher to ask for more novels, her health markedly declined, and some hinted at senility (see Morgan, 1984, pp. 370–71 and Lancashire 2010), although she was never diagnosed with AD.

1.4 Our goals and hypotheses

Because we are looking at published text, we are looking for early signs of dementia (when publishable writing is still possible). We focused, then, on lexical and syntactic markers: vocabulary size, repetition, word specificity, word-class deficit, fillers, grammatical complexity, and the use of passive. We expect that Murdoch's data will exhibit the linguistic patterns of dementia patients and P.D. James's data the linguistic patterns of healthy aging described in Section 1.2 above, as summarized in Table 1. Further, we hypothesize that Agatha Christie's data will show patterns of decline similar to those of AD patients.

To avoid the problems evident in Garrard *et al.*'s study, we used far larger corpora—the complete text of fifteen to twenty novels by each author—with a complete syntactic analysis of each text and more-meaningful measures of lexical and syntactic complexity.

2 Materials

2.1 Data

We analyzed twenty of Murdoch's twenty-six novels, published between ages 35 and 76 ($M = 52.7$), sixteen of Christie's novels written between ages 28 and 82 ($M = 59.0$), and fifteen of the novels of P.D. James, published between ages 42 and 82 ($M = 63.9$). The Appendix lists the fifty-one novels, with publication years and estimated ages of the authors at composition. Apart from Christie's *Curtain*, written in the early 1940s but published only in the mid-1970s, we assume, given no

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

Linguistic marker	Normal aging	Dementia
Lexical		
Vocabulary size	Gradual increase, possible slight decrease in later years	Sharp decrease
Repetition	Possible slight decrease/increase	Pronounced increase
Word specificity	Possible slight increase/decrease	Pronounced decrease
Word class deficit	Insignificant change	Pronounced deficit in nouns; possible compensation in verbs
Fillers	Possible slight increase	Pronounced increase
Syntactic		
Overall complexity	No change or gradual decline, possible rapid decline around mid-70s	Sharp decline
Use of passive	Possible slight decrease	Pronounced decrease
Auxiliary verb	*Be*-passives dominate	*Get*-passives dominate
Agentless passive	Moderate decrease	Greater decrease

evidence to the contrary, that each novel was written just prior to the year of its publication. All texts belong to the same genre, prose fiction, and in all cases, the novels span the author's career.

The text of two of Christie's novels, *The Mysterious Affair at Styles* (1920) and *Secret Adversary* (1922), came from Project Gutenberg; all the others were scanned and converted to plain text with commercial optical character recognition (OCR) software (the Christie novels with OmniPage Professional 15.0, and the others with ABBYY FineReader 9.0 Professional Edition). We corrected OCR errors in spelling and punctuation manually, and then common error patterns semi-automatically with an interactive script.

The experiments to be described below require various levels of processing of the resulting text files, ranging from simple unlemmatized word sequences through lemmatized sequences and part-of-speech tagging to complete syntactic analyses. Given the plain text file of a novel, we first separated punctuation marks and clitics from the word tokens to which they are attached (e.g. *I 'm, is n't, John 's*), and lemmatized the words with NLTK WordNet's *morphy* method (Bird *et al.*, 2009). We then determined sentence boundaries with a rule-based, deterministic algorithm. Then a parse tree was generated for each sentence, using the Charniak (2006) parser, which includes part-of-speech tagging as a subprocess. Finally, a script was run on the parse trees to correct common patterns of error made by the parser. For detailed descriptions of the implementation, see Le (2010).

2.2 Impact of writing process and genre on reliability

Our experiments assume that the text we are analyzing is solely that of the author whom we want to study, without any influence from a collaborating writer or any significant alteration by an editor. In addition, the process by which the author composed might affect the linguistic markers being analyzed.

Agatha Christie's writing process changed over time. She wrote her earlier novels in longhand and then typed them on a typewriter (after hiring a secretary to take dictation for a brief period), but, on breaking her wrist in 1952, she began using a Dictaphone. She wrote in her autobiography that subsequent removal of repetitions made during the recording was "irritating" since it "destroys the smooth flow which one gets otherwise" (Christie, 1977, p. 348). Her editing routine might have changed again at the time she composed her second-last novel, *Elephants can Remember*, when she broke her hip, and for her final novel, *Postern of Fate*, her agent requested that Christie receive editing help (Morgan, 1984, p. 371).

All the novels that Christie wrote under her own name[4] are of the detective genre, except for the political spy thriller *Passenger to Frankfurt: An Extravaganza* (1970). Lancashire (2010; Lancashire

and Hirst, 2009) observed that this novel has a far larger vocabulary than her detective novels, owing to the extensive research into relevant political literature that Christie carried out while composing it, thus enriching it with a vocabulary beyond her own. Lancashire and Hirst (2009) considered this novel an outlier (since it departed from the detective-novel genre) and so excluded it from the dataset they used to measure vocabulary. We follow the same view here only with regard to vocabulary; there is no *a priori* reason to consider this novel an outlier with regard to syntax.

Iris Murdoch wrote in longhand, without a typewriter or word processor, after months of working out the plot, 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). She denied anyone else the right to edit her writings, even her publishers.

P.D. James begins with months of ''plotting and planning'' sketched in a notebook, penning out-of-order sequences by hand, putting the book together, and then dictating it to a secretary (James, 2010). Little is known about her editorial process, apart from her use of a dictionary and a thesaurus. However, James commented that she found the writing process of her latest novel easy, and there is no reason to believe that she relied unusually on reference sources or that her writing is significantly modified by an editor.

To the best of our knowledge, except as noted above, no novel in our dataset departs from the usual writing methodology of its author, belongs to an atypical genre, or involves research to the degree that it should be judged an outlier.

3 Lexical analysis

In this section, we present five analyses of the novels at the lexical level, using a variety of measures that will be explained in the sections below. To look for changes over time in each of the measures, we performed simple linear regression of the measure against the author's age, and tested each regression model for a statistically significant relationship between the author's age and the value of the measure. We tested correlation between measures with the Spearman rank-order correlation coefficient method.

Some of the measures that we used are sensitive to text length and therefore required a cut-off threshold. For example, in the case of vocabulary richness, a 60,000-word novel will not have twice as many unique word-types as a 30,000-word novella, since the number of word-types does not grow linearly with the number of word-tokens (the second half of the 60,000-word novel is bound to ''reuse'' many word-types of the first half). As each text in our dataset contains at least 55,000 tokens, except for Murdoch's *The Italian Girl* with 48,448 tokens, for length-sensitive measures, we considered only the first 55,000 tokens of each text, and excluded *The Italian Girl* to avoid lowering the token count. For the remaining measures, we used the complete text of all the novels. In our graphs, the outlier in the Christie dataset, *Passenger to Frankfurt,* is excluded from the overall trend but included as a single datapoint.

3.1 Vocabulary size

We measured the vocabulary size of each novel by the type/token ratio (TTR), i.e. the number of unique lemmatized word-types divided by the total number of word-tokens, and by the word-type introduction rate (WTIR), i.e. the cumulative number of unique lemmatized types computed at every 10,000-token interval.

Figure 1 displays the TTR variations of each author over time. Murdoch's ratio drops at the age of 50, and then begins to rise and peak in her mid-60s at *The Sea, the Sea,* before plummeting with her last novel, *Jackson's Dilemma*, written in her 75th year. Although the dataset as a whole does not exhibit a statistically significant change [$F(1,19) = 0.19$, $P = 0.6651$], 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 five novels, the decreasing trend is steeper and also significant [$F(1,3) = 14.17$, $P = 0.0328$]. Christie's ratios vary between 0.07 and 0.084 before her 60s, and then begin to drop, reaching bottom at her second last novel, *Elephants Can Remember.*

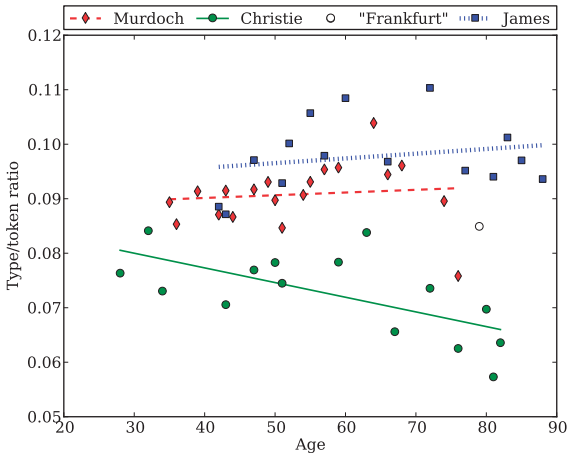

Fig. 1 Type/token ratio within the first 55,000 tokens

A significant decline [$F(1,13) = 9.29$, $P = 0.0093$] is found for the entire period, excluding *Passenger to Frankfurt,* whose TTR is higher than Christie's other novels. P.D. James's TTR varies in the 0.09 to 0.11 range with no signs of decline; the slight rising trend is statistically insignificant [$F(1,13) = 0.59$, $P = 0.4550$].

Word-type introduction rates appear in Figs 2 and 3. Each line reflects the vocabulary growth of one novel measured at every 10,000-word-token interval. Lines may overlap, and their clustering indicates some consistency. Each author's novels are divided into two groups, earlier and later, the former represented as dotted lines and the latter as solid lines. For Christie, this division coincides with her change of writing method, from the typewriter to the Dictaphone. Her novels, containing from 55,000 to under 80,000 word tokens, are measured up to a maximum of 70,000 tokens. For a fair comparison, the graphs for Murdoch and James in Fig. 2 are scaled to focus on the first 70,000 tokens; the graphs for the complete novels are shown in Fig. 3.

Murdoch's last novel, *Jackson's Dilemma,* has an unusually low rate of vocabulary growth compared to her previous works, all of which (with the exception of *The Philosopher's Pupil*) cluster together in a concentrated band. This confirms the TTR results that the decline in Murdoch's vocabulary occurred abruptly, and is consistent with Garrard *et al.* (2005)

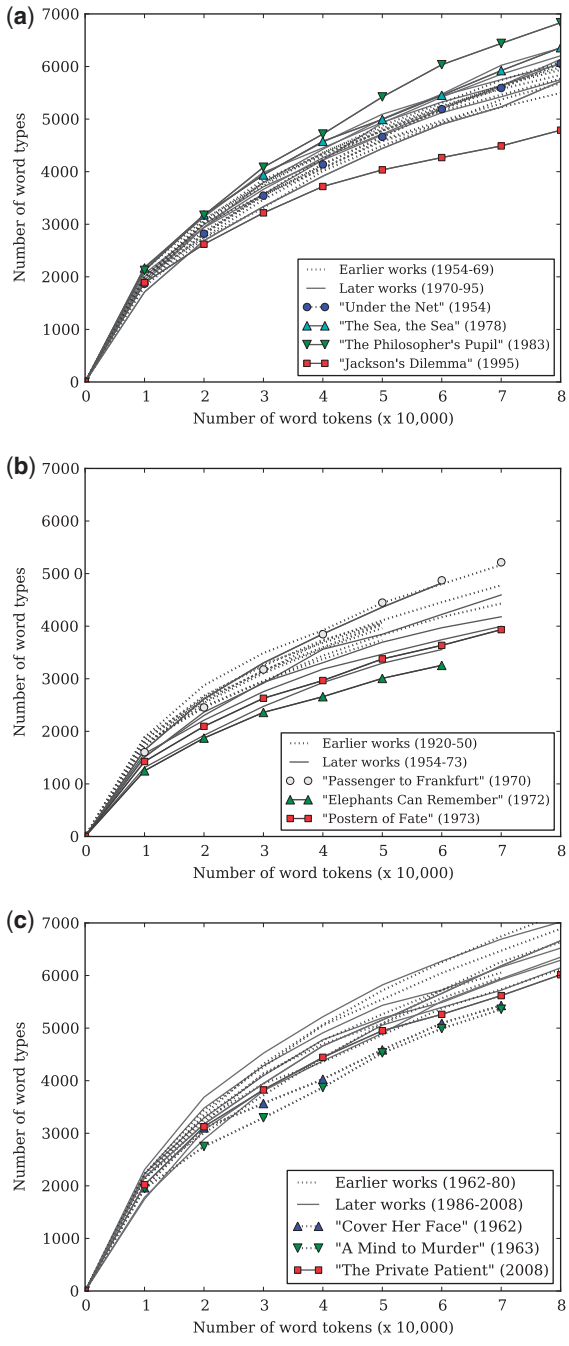

Fig. 2 Word-type introduction rate up to the 70,000th token. (**a**) Iris Murdoch (**b**) Agatha Christie (**c**) P.D. James.

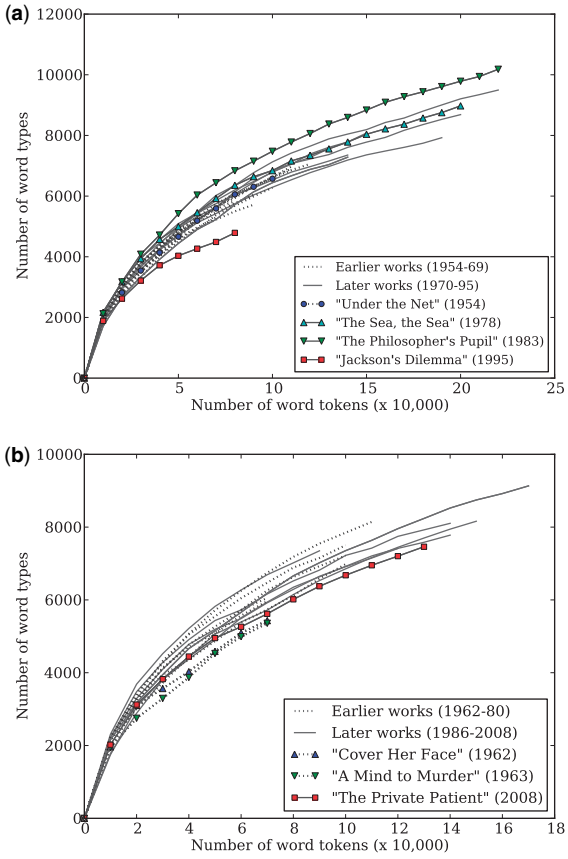

Fig. 3 Word-type introduction rate (complete texts). (**a**) Iris Murdoch (**b**) P.D. James.

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; Fig. 3a shows that the vocabulary growth of *Jackson's Dilemma* slows down significantly only *after* the 40,000th token, as distinguished from most of Murdoch's works.

A more gradual declining tendency appears 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 her 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 are relatively consistent,

Table 2 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 $P < 0.01$).

with earlier and later works intertwined, apart from her first two novels, which stand out in a slightly lower range. James's most recent novel remains in the mid-range up to the 50,000-token mark, then converges towards her lower range from the 60,000-token mark onwards, but does not greatly depart from her usual rate.

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.

Table 2 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.

3.2 Lexical repetition

While an author sometimes uses deliberate repetition for effect, an increasing rate of lexical repetition may indicate a reduced vocabulary or word-retrieval difficulties. We performed two analyses: global and local.

The first measured an author's tendency to repeat by counting *global* word *n*-gram repetitions, that is, phrases containing from two to eleven words that occur at least twice at any point in a text. (We are using the word *phrase* here to refer to word *n*-grams, not syntactic constituents.) We extracted all fixed phrasal repetitions, as defined by length in

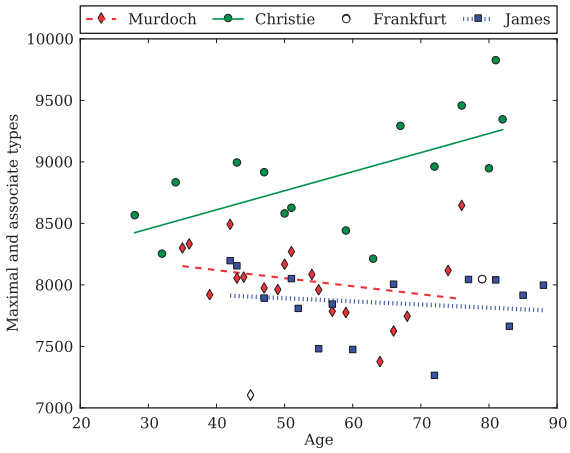

Fig. 4 Maximal and associate phrasal repetitions (types)

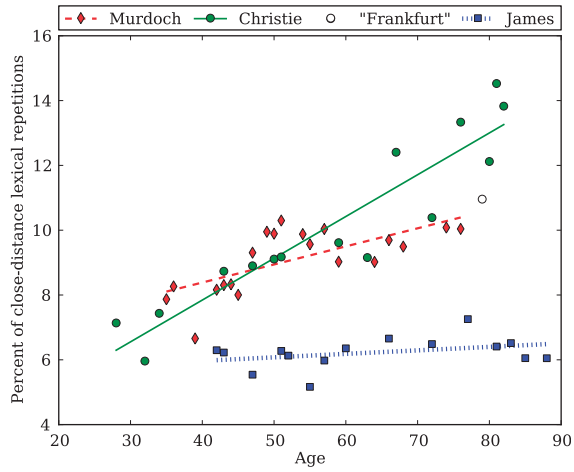

Fig. 5 Lexical repetitions within 10 subsequent content words

words and frequency (Lancashire *et al.*, 1996, p. 97) in the first 55,000 tokens of each novel. We define *maximals* to be the longest repeating fixed phrases in a text that are not found inside any other repeating fixed phrase, and *associates* to be substrings of maximals that occur more frequently than those maximals. Figure 4 shows that Christie's unique repeated phrases of both types, maximals, and associates, rose by about 800 over her career (8,567–9,346), a significant increase [$F(1,13) = 8.53$, $P = 0.0119$] that peaked at 9,827 in *Elephants*. Although Murdoch's repeating phrases, like James's, fell overall, but not significantly [$F(1,17) = 1.27$, $P = 0.2760$], repetitions rose again in the last five Murdoch novels, peaking in *Jackson's Dilemma*.

The second analysis measured *local* repetition: the proportion of lemmatized open-class words (i.e. nouns, content verbs, adjectives, and adverbs) repeated within 10 subsequent open-class words, computed over the number of all content words in each novel. Figure 5 shows the results. Murdoch's overall repetition rate increases significantly, peaking in her 51st year, and again rising in her last two novels. Christie's repetition rates rise more steeply, with high certainty, peaking in her last two novels, of which 14.53% and 13.83% of the content word-tokens are repeated within close distance; this stands in sharp contrast to the rates of 7.14% and 5.96% in her first two novels. Repetition rates in James's novels remain relatively stable in the low range (5.54–7.26%). When the distance is extended to 20 subsequent content words, similar patterns are observed in all three authors, as indicated by the coefficient values in Table 3. The rise is pronounced in Christie, slightly less steep in Murdoch—both with statistical significance—and marginal without significance in James.

Table 4 displays the correlation between rates of lexical repetition of distance 20 and the other vocabulary measures, PR (phrasal repetition), 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. 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 are also in a trough then.

3.3 Lexical specificity

We approximate lexical specificity by computing the proportions of indefinite nouns and of high-frequency, low-imageability verb tokens in

Table 3 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

*$P < 0.05$; **$P < 0.01$.

Table 4 Correlation between repetition measures and vocabulary measures

	PR	TTR	WTIR 50,000
Murdoch			
LR	+0.17	−0.14	−0.11
PR	–	−0.93**	−0.91**
Christie			
LR	+0.82**	−0.79**	−0.79**
PR	–	−0.95**	−0.95**
James			
LR	+0.58*	−0.52*	−0.45
PR	–	−0.97**	−0.94**

*$P < 0.05$; **$P < 0.01$. LR, lexical repetition within 20 tokens; PR, phrasal repetition (global); TTR, type/token ratio; WTIR 50,000, word-type introduction rate at 50,000 tokens.

each text. A higher proportion indicates greater reliance on generic words and, consequently, a lower overall specificity rank. We considered four indefinite nouns—*thing(s), something, anything, nothing*—and 35 high-frequency verbs of relatively low specificity, in their base and conjugated forms [the first 14 of which are present in the writing samples of semantic dementia patients (Bird *et al.*, 2000)]:

> *be, come, do, get, give, go, have, know, look, make, see, tell, think, want, ask, feel, find, forget, happen, hear, like, live, mean, meet, put, remember, run, say, seem, speak, suppose, take, use, walk, wonder.*

Figure 6, which displays the number of indefinite nouns, shows a slight and gradual increase for most of Christie's writing career, with an abrupt climb to a peak in her last two novels, constituting a highly significant trend [$F(1,13) = 14.44$, $P = 0.0022$]. Both Murdoch's and James's usage of indefinite nouns

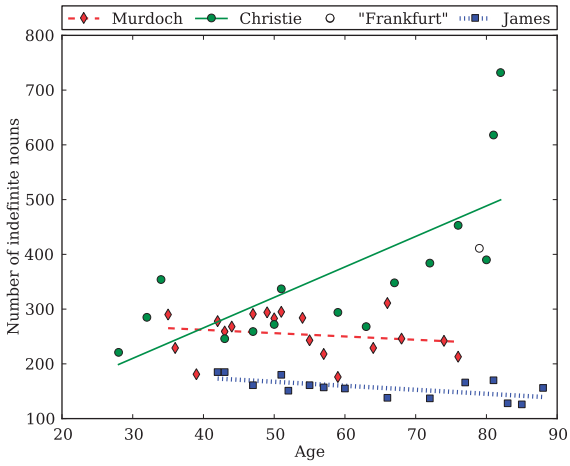

Fig. 6 Number of indefinite noun occurrences

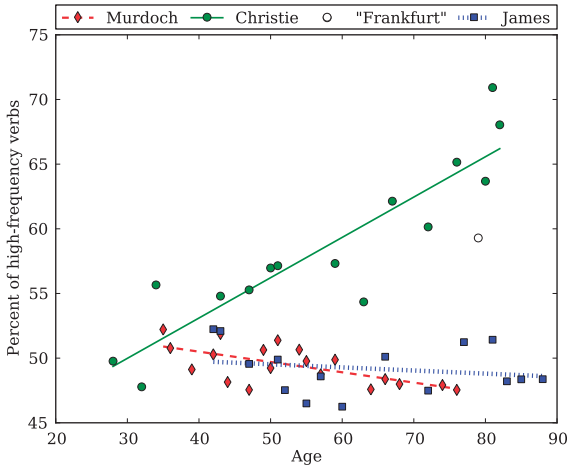

Fig. 7 Proportion of 35 high-frequency verbs

decreases over time, although James's rates are more consistent and hence statistically significant [$F(1,13) = 8.07$, $P = 0.0139$], while Murdoch's fluctuate [$F(1,17) = 0.63$, $P = 0.4368$].

Figure 7 displays the percentages of the indefinite verbs in the first 55,000 tokens of each novel. Christie's rates show a marked increase, from a low of 48% for *The Secret Adversary* (1922), to a high of 71% for *Elephants Can Remember* (1971) [$F(1,13) = 55.74$, $P < 0.0001$]. That thirty-five verb-types account for 71% of all verb-tokens 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 is found [$F(1,17) = 12.13$, $P < 0.0028$], while James's slight decreasing trend is statistically insignificant [$F(1,13) = 0.53$, $P = 0.4789$].

Table 5 shows these approximations to be significantly correlated with the lexical repetition measure and negatively correlated with the vocabulary measures for both Christie and James, showing that a larger vocabulary entails fewer lexical repetitions and less reliance on common verbs of low specificity.

Table 5 Correlation between specificity approximations and other lexical measures

	IN	LR	PR	TTR	WTIR 50,000
Murdoch					
HFV	+0.29	−0.11	+0.32	−0.24	−0.28
IN	–	+0.19	+0.14	−0.15	−0.18
Christie					
HFV	+0.88**	+0.96**	+0.78**	−0.79**	−0.79**
IN	–	+0.78**	+0.65**	−0.70**	−0.70**
James					
HFV	+0.68**	+0.60*	+0.93**	−0.88**	−0.90**
IN	–	+0.17	+0.70**	−0.68**	−0.79**

*$P < 0.05$; **$P < 0.01$. IN, indefinite nouns; HFV, high-frequency verbs; LR, lexical repetition within 20 tokens; PR, phrasal repetition (global); TTR, type/token ratio; WTIR 50,000, word-type introduction rate at 50,000 tokens.

3.4 Word-class deficit

In this section, we look at the proportions of each word class over the entire length of each text, in terms of both word-tokens, in order to look for signs of deficit in or reliance on individual classes, and word-types, in order to measure vocabulary size of open classes.

Figures 8–12 display the changes in the proportions of nouns, pronouns, content verbs, adjectives, and adverbs, in terms of token count and type count. Table 6 shows the results of statistical significance tests for each word class of interest, and Tables 7 and 8 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 full, context-aware part-of-speech tagging, discovers longitudinal variations in the datapoints. Among the important findings are a decline in noun-token proportion and a rise in verb-token proportion in both Murdoch's and Christie's novels (Figs 8a and 9a).

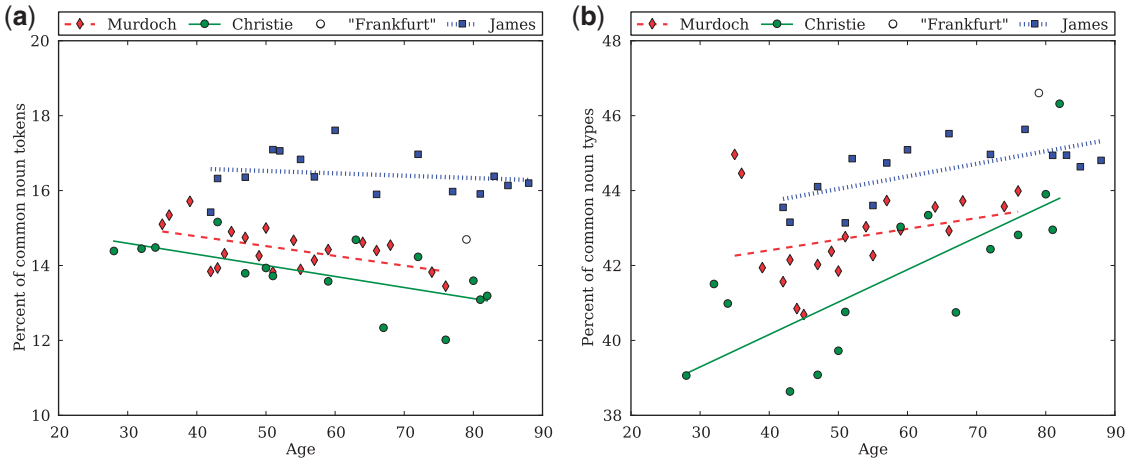

Fig. 8 Proportion of common nouns: (**a**) by token (**b**) by type

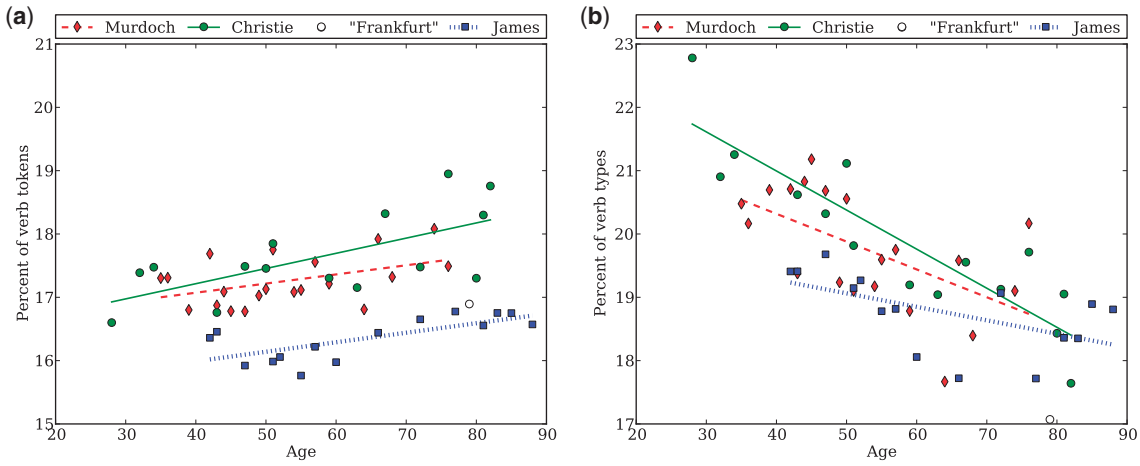

Fig. 9 Proportion of content verbs: (**a**) by token (**b**) by type

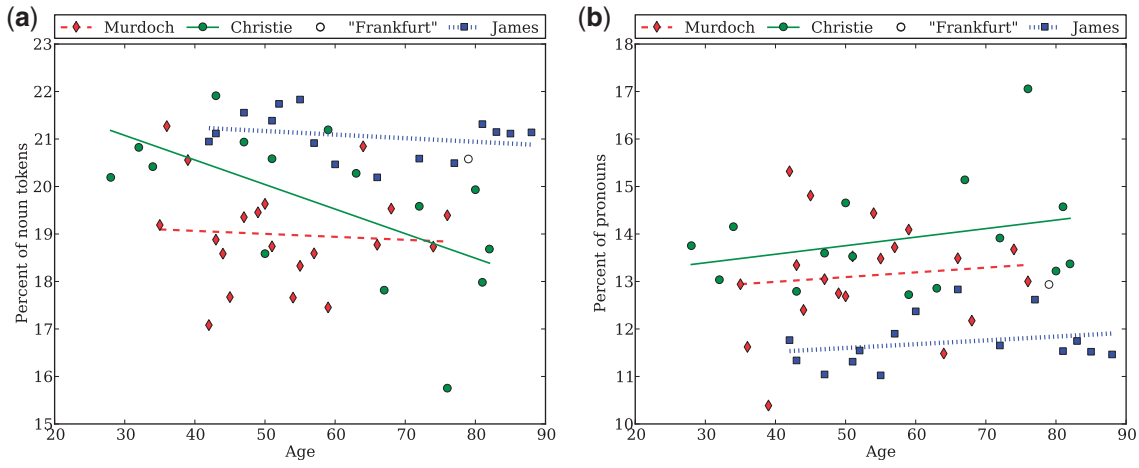

Fig. 10 Proportion of nouns and pronouns: (**a**) common and proper nouns (**b**) pronouns

These trends are statistically significant, with *P*-value between 0.0076 and 0.0469 (Table 6). Our statistical tests also showed a negative correlation between the noun and verb proportions of the two authors, which is stronger and more significant in Christie's results (Table 7). Similar correlations occur in semantic dementia patients (Bird *et al.*, 2000) in that the apparent noun deficit is compensated for by a rise in verbs. On the other hand, no significant trend is found in James's noun proportion and, although verb proportion shows a slight increasing tendency with high significance, the two values are not highly correlated. Nor are the proportions of pronouns and that of common nouns for all three authors, as shown in Table 7.

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 (Table 7). Figure 8 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 6 shows, a significant decreasing tendency is found

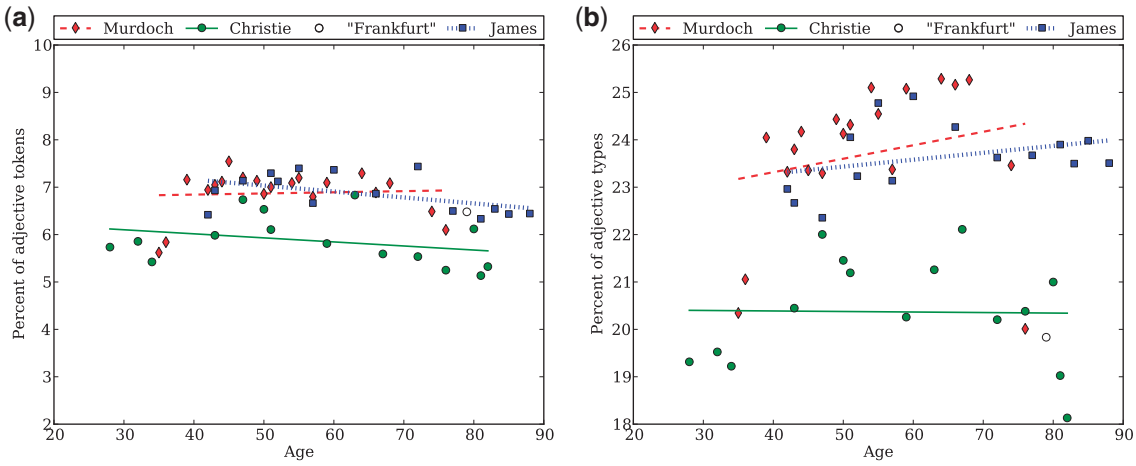

Fig. 11 Proportion of adjectives: (**a**) by token (**b**) by type

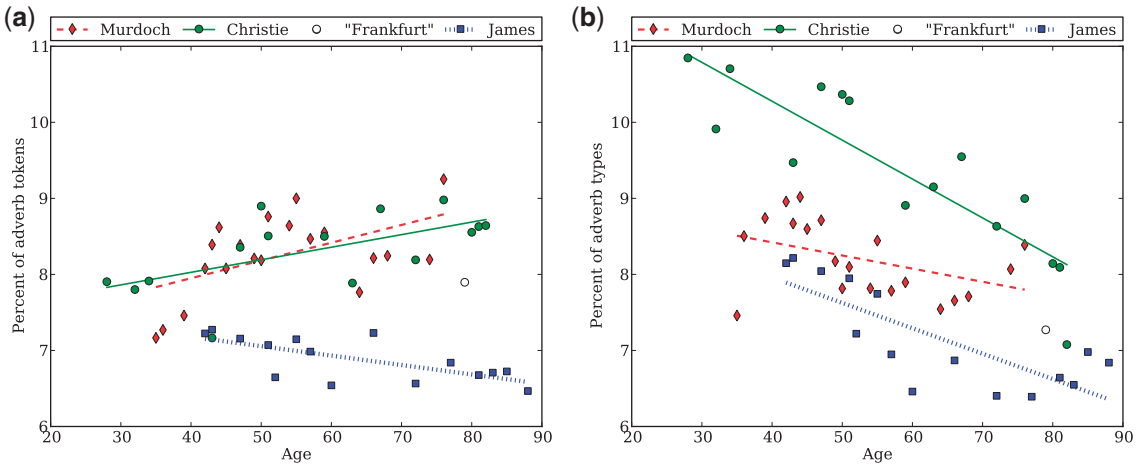

Fig. 12 Proportion of adverbs: (**a**) by token (**b**) by type

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 nouns is remedied by increased use of pronouns, in addition to the previously mentioned rise in verb proportion.

An opposite tendency is observed when types are considered instead of tokens (Figs 8b and 9b). Noun proportions increase while verb proportions decrease for all three authors in varying degrees. These trends are all significant, except for Murdoch's noun proportion (Table 6). 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).

Table 6 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 7 Correlation between word class proportions (in tokens)

	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							–

*$P < 0.05$; **$P < 0.01$.

Similarly, a disconnection between type and token exists in the proportions of adjectives and adverbs (Figs 11 and 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.

Table 8 Correlation between word class proportions (in types)

	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				–

*$P < 0.05$; **$P < 0.01$.

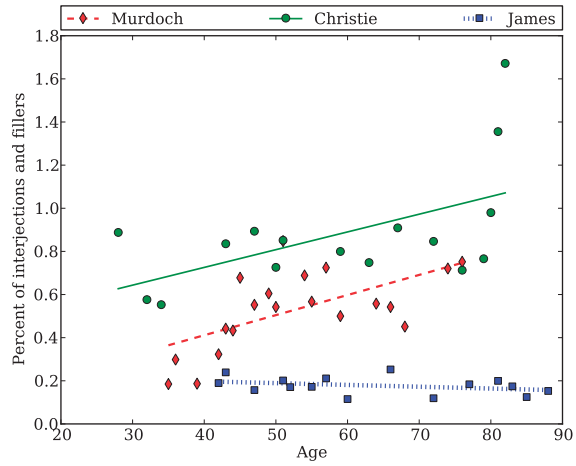

Fig. 13 Proportion of interjections and fillers

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 6, is steepest and highly significant for Christie, moderate and significant for James, and slightest and approaching significance for Murdoch.

Tables 7 and 8 report the correlation coefficients between different word classes in token and in type. The rise of verb-token proportion in Christie's novels is positively correlated with the rise of adverb-token proportion, while this is not the case for Murdoch and James. In light of our high-frequency verb results, which show 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.

3.5 Fillers

Our final lexical measure is the proportion of words identified in part-of-speech tagging as interjections and fillers. While these words largely appear in 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. Nonetheless, this measure may reflect an author's stylistic choice rather than a cognitive decline and therefore should be interpreted cautiously.

The proportions of lexical fillers and interjections are shown in Fig. 13. Consistent with our prediction, Murdoch's and Christie's data indicates 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 seen in her 30s (0.55). James's data, on the contrary, remains consistently low throughout, following a slight decreasing trend that is not statistically significant [$F(1,13) = 1.60$, $P = 0.2282$].

As Table 9 shows, this measure is moderately correlated, with statistical 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), phrasal repetitions (PR), indefinite nouns (IN), and high-frequency verb proportions (HFV). These results make intuitive sense: a high rate of fillers may indicate word-finding difficulty, which reduces vocabulary size, increases repetitions, and leads to a

Table 9 Correlation between proportion of fillers and other lexical measures

	TTR	WTIR 50,000	LR	PR	HFV	IN
Murdoch	+0.04	+0.07	+0.80**	−0.04	−0.30	+0.08
Christie	−0.50	−0.50	+0.63*	+0.50	+0.53*	+0.30
James	−0.59*	−0.61*	+0.48	+0.65**	+0.72**	+0.51

*$P < 0.05$; **$P < 0.01$. TTR, type/token ratio; WTIR 50,000, word-type introduction rate at 50,000 tokens; LR, lexical repetition within 20 tokens; PR, phrasal repetition (global); HFV, high-frequency verbs; IN, indefinite nouns.

greater reliance on generic verbs. For James, a similar correlation is found between this measure and most other measures, except for lexical repetition, whereas for Murdoch the situation is reversed. Because of the varying degrees of correlation, we neither reject these results nor use them as a basis for our conclusion.

4 Syntactic analysis

Most of our syntactic measures operate on parse trees, one for each sentence in a text, generated by the Charniak parser. As in the lexical analysis, we performed simple linear regression on each set of results and tested them for statistical significance. Correlation between measures was also computed with the Spearman correlation coefficient method. None of the novels was excluded or considered an outlier.

4.1 Syntactic complexity

Syntactic complexity is assessed by several measures, described below, that have been shown to be sensitive to the effects of aging (Cheung and Kemper, 1992). However, the dialogues in novels may complicate analysis of syntactic complexity. Fiction writers often try to capture the essence of natural, real-life conversations in their dialogues. Because 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 may partly determine its complexity scores. Ideally, we would perform separate syntactic analysis on the dialogue portions and the narrative portions; however, this separation of dialogue from narrative cannot be accomplished, given the limitations of our scanned texts (see Le, 2010 for a detailed discussion of the problems). Consequently, the results of these measures should be interpreted cautiously.

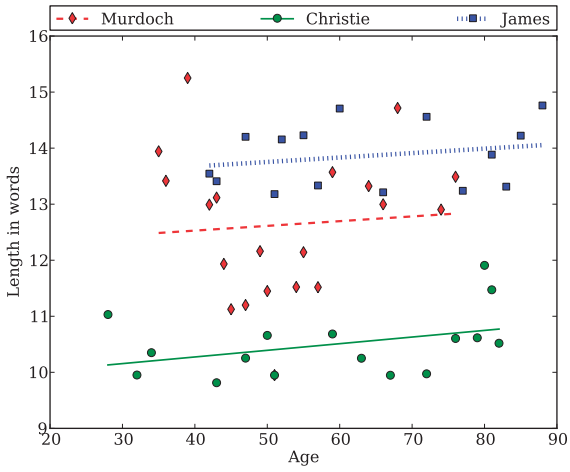

Fig. 14 Mean length in words per sentence

4.1.1 *Mean length of utterance and mean number of clauses per utterance*

For each sentence parse tree, the number of words and the number of clauses (main, subordinate, and embedded) are counted. (Contractions, such as *is n't* and *they 're,* count as two words.) Mean length of utterance (MLU) and mean number of clauses per utterance (MCU) are the respective averages over all sentences in a text.

The MLU results of all three authors show increasing tendencies, but none are statistically significant. Murdoch's mean sentence length reaches a peak early in her career, plummets to a low at age 51 years with *Defeat* (which overlaps with Christie's datapoint for *Towards Zero* at age 51 years in Fig. 14—the two points are indistinguishable in the figure), 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 years, and 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 most recent work.

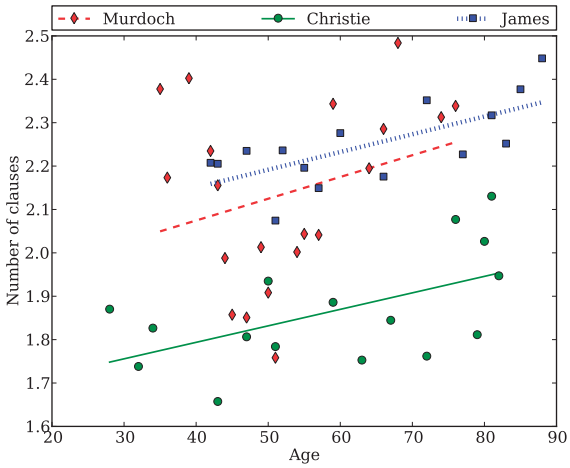

Fig. 15 Mean number of clauses per sentence

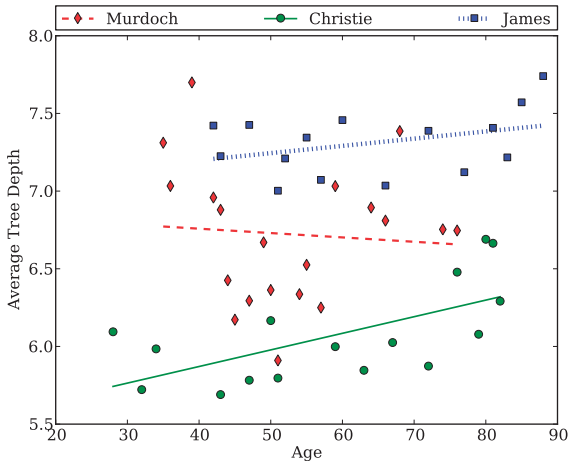

Fig. 16 Average parse tree depth

The MCU data (Fig. 15) show a significant overall increasing trend 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 most recent novel. Murdoch's overall trend is insignificant because of an unexpected 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*.

4.1.2 *Parse tree depth and Yngve depths*

The parse tree depth measure computes the average maximum depths of the parse trees of the sentences in each complete novel. This reflects the average number of embedded structures in a sentence, in order to approximate syntactic complexity, relying on the assumption that deeply nested levels of embedding are associated with complex sentences. 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.

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 complex and put a heavier requirement on working memory (Kemper *et al.*, 2001). Therefore, we also used an asymmetric measure that compensates for left-branching structures, the Yngve (1960) measure, which assigns a higher score for left-branching syntax than for right-branching. However, the results of this measure were essentially the same as those of the regular depth measure previously described; see Le (2010) for details.

4.1.3 *D-Level*

Originally constructed by Rosenberg and Abbeduto (1987), the D-Level scale is a psycholinguistics-based ranking of sentence types into eight levels of increasing syntactic complexity. Cheung and Kemper (1992) and Covington et al. (2006) addressed some problems in the original scale and proposed modifications to better model incremental levels of complexity. Our implementation, based on the revised version of D-Level by Covington et al.

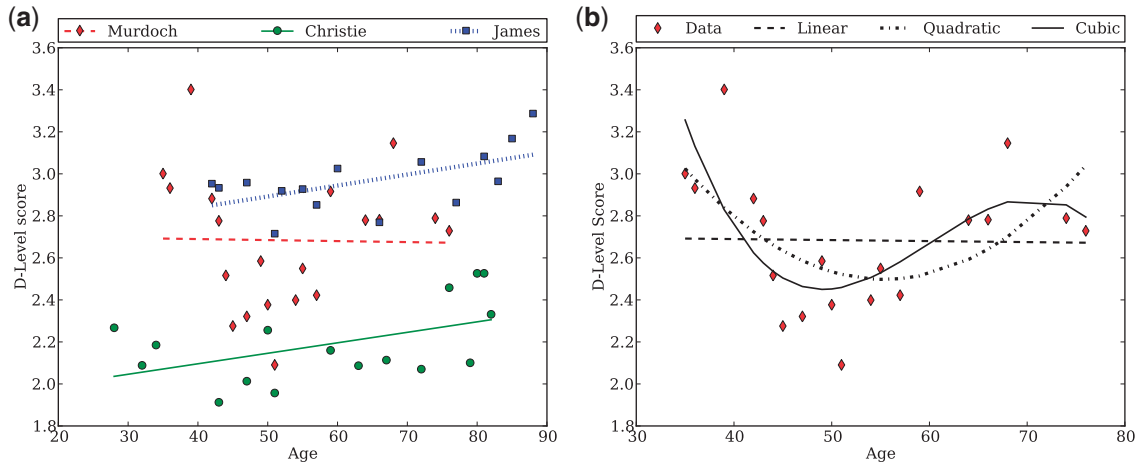

Fig. 17 Average D-level score: (**a**) All three authors (**b**) Murdoch

Table 10 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
Max.Yngve	0.0015	0.17	0.0029	8.70*	0.0005	0.16
Total Yngve	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 11 Correlation between syntactic complexity measures

	1	2	3	4	5
Murdoch					
1. MCU	–	+0.94**	+0.92**	+0.97**	+0.93**
2. MLU		–	+0.98**	+0.97**	+0.94**
3. Max.Yngve			–	+0.96**	+0.91**
4. Total Yngve				–	+0.90**
5. D-Level					–
Christie					
1. MCU	–	+0.78**	+0.67**	+0.78**	+0.92**
2. MLU		–	+0.76**	+0.95**	+0.80**
3. Max.Yngve			–	+0.86**	+0.73**
4. Total Yngve				–	+0.83**
5. D-Level					–
James					
1. MCU	–	+0.70**	+0.68**	+0.65**	+0.93**
2. MLU		–	+0.93**	+0.95**	+0.75**
3. Max.Yngve			–	+0.94**	+0.67**
4. Total Yngve				–	+0.63*
5. D-Level					–

*$P < 0.05$; **$P < 0.01$.

(2006; and detailed in Le 2010), uses pattern-matching to determine whether a parse tree matches the constructions indicative of each level. Each parse tree is given a score between 0 and 7, and these scores are averaged over the entire novel.

Figure 17 displays results of the D-Level measure, which are largely similar to the results of other syntactic measures. Murdoch's average D-Level scores over time exhibit a very slight, statistically insignificant decrease. 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 [$F(1,13) = 6.33$, $P = 0.0258$].

As summarized in Table 10, 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, rather than linear and quadratic models, as demonstrated in Fig. 17b.

Table 11 shows the correlation coefficients between the complexity measures, which are mostly

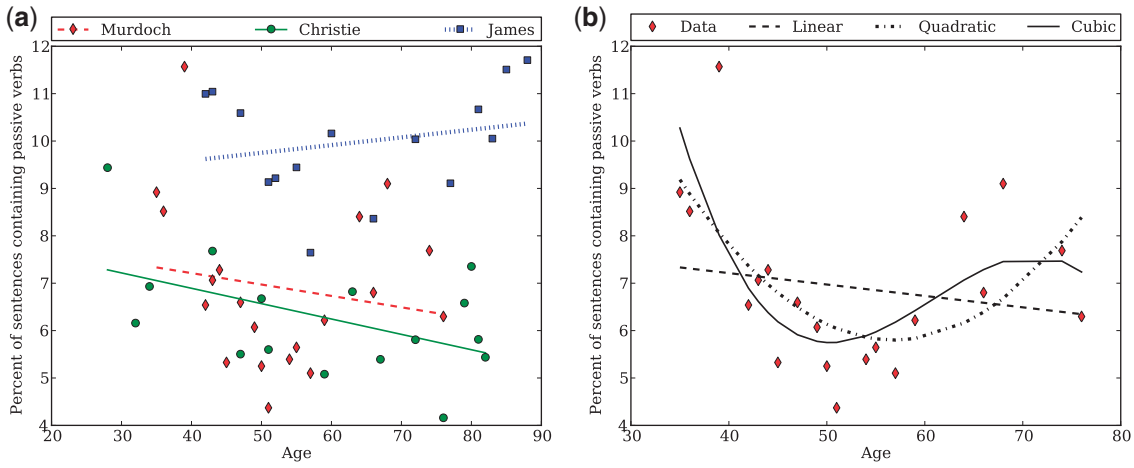

Fig. 18 Proportion of passive sentences: (**a**) All three authors (**b**) Murdoch

Table 12 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

*$P < 0.05$; **$P < 0.01$.

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 in her later novels.

4.2 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. The same pattern-matching algorithm used for the D-Level measure identifies the three passive forms (see Le, 2010 for details). The measure reports the percentage of sentences containing passive forms over the total number of sentences, as well as the percentages of each passive form over all passive sentences. Note that, because 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%.

Figure 18a 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 12. Christie's decline, however, approaches significance with a *P*-value of 0.0541.

Table 13 Correlation between passive proportion and syntactic complexity measures

	MCU	MLU	Max.Yngve	Total Yngve	D-Level
Murdoch	+0.69**	+0.77**	+0.70**	+0.69**	+0.81**
Christie	−0.27	+0.14	+0.19	+0.06	−0.04
James	+0.65**	+0.56*	+0.20	+0.18	+0.81**

*$P < 0.05$; **$P < 0.01$.

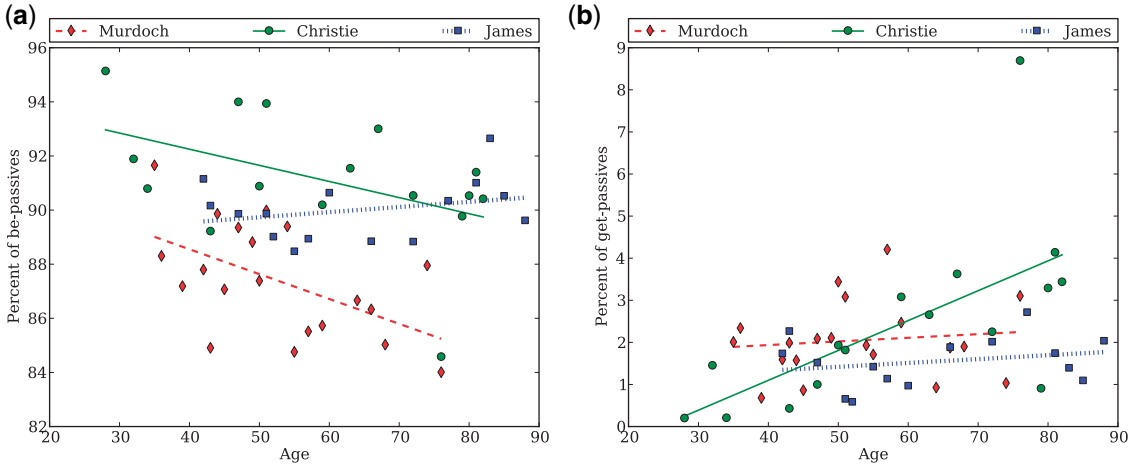

Fig. 19 Proportions of *be*-passives and *get*-passives: (**a**) *be*-passives (**b**) *get*-passives

Table 13 shows that this measure is moderately correlated with most syntactic complexity measures for Murdoch and James, but clearly not for Christie. This suggests that access to passive forms may be 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 18b, which is consistent with the previous observation of a syntactic decline in her 50s.

Figure 19a and b 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 statistically 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, whereas Christie's results, in contrast, follow a strong, highly significant rising pattern, peaking abruptly in her 1967 novel, *Endless Night*. Proportions of passives with *by*-phrase, shown in Figure 20, 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 are given in Table 12.

5 Discussion

In this section, we discuss our results with respect to the hypotheses of Table 1. A summary is shown in Table 14.

5.1 Lexical analysis

Our lexical analysis results largely support our hypotheses and the expected lexical patterns of linguistic change given in Table 1. TTR and WTIR in Murdoch's later novels indicate her lexical decline,

especially in *Jackson's Dilemma*, with an abrupt drop in vocabulary size in the second 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 a compensatory increase in verb-token proportion. Murdoch's repeating phrases decline until she is in her mid-60s, when they rise steadily and peak with *Jackson's Dilemma*. Unexpectedly, in 1969–70, more than twenty years before Alzheimer's symptoms became clear, her vocabulary declined (see section 5.3 below). The first signs of her final vocabulary decline appear in *The Green Knight* (1993), composed ca. 1992, five years before her diagnosis of AD, and seven years before her death. But contrary to our prediction, her lexical specificity remained intact.

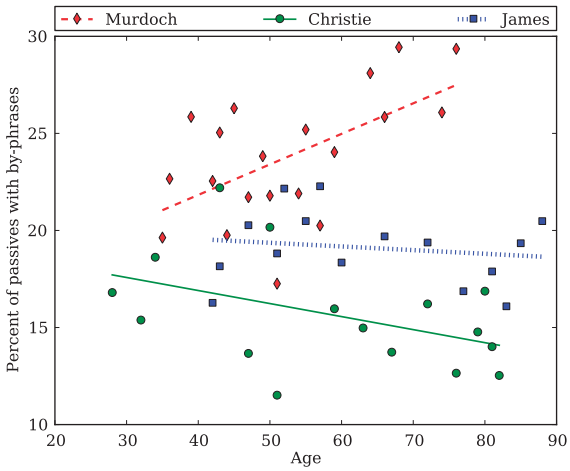

Fig. 20 Proportion of passive sentences with *by*-phrase

The lexical results of James's novels follow the patterns expected for normal aging elders. Her vocabulary, repetition, and specificity scores vary only in a relatively small range, with no apparent word-class deficit. Vocabulary growth of the latter half of James's most recent book, *The Private Patient*, tapers slightly but does not stray far from her average range.

Table 14 Patterns of linguistic changes observed in the novels of Murdoch, Christie, and James

Linguistic marker	Murdoch	Christie	James
Lexical			
✓ 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)
✗ Phrasal repetition	(Decrease); rise in last five novels	Increase	(Decrease)
✓ Word repetition	Strong increase overall; sharp rise in her 50s	Pronounced increase	(Marginal increase)
✗ Noun specificity	(Slight decrease)	Pronounced increase	Slight decrease
✗ Verb specificity	Moderate decrease; noticeable rise in her 50s	Pronounced increase	(Slight decrease)
✓ 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)
✓ Fillers*	Pronounced increase overall; noticeable rise in her 50s	Pronounced increase	(Slight decrease)
Syntactic			
✗ Overall complexity	Irregular changes; deep decline in her 50s	(Minor changes)	(Minor changes)
✗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)

*May also reflect an author's stylistic choices in creating natural dialogues. The items reported in parentheses are statistically insignificant trends. Check-marks indicate that the patterns observed follow our hypotheses; crosses indicate otherwise.

Christie's vocabulary, repetition, and specificity measures show an overall decline with just two exceptions, *Destination Unknown* (1954, at age 63 years) and the known outlier, *Passenger to Frankfurt* (1970, at age 79 years). The latter exhibits the largest vocabulary of the 16 Christie novels we analyzed, the highest word-type introduction rate, and 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 had 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*. (Christie complained about the repetitions endemic in dictating, and possibly her illness interfered with her routine pruning of them.) Consistent with our predictions, we find a deficit in noun tokens that is significantly correlated with a 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 significantly and, since it is highly correlated with the rise in verbs, may have been a compensation for the loss of specific, descriptive verbs.

Although Murdoch and Christie exhibit somewhat different linguistic behaviour (the latter's dramatic rise in indefinite nouns is not found in Murdoch), their common lexical decline—in contrast to James—validates our hypotheses with respect to lexical markers.

5.2 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 unexpectedly reveal an abrupt drop in her late 40s and 50s, and then a period of recovery which, for some measures, is followed by a slight decline in her last two novels. This early syntactic drop occurs in all novels that we analyzed, parallels a trough in her vocabulary, and so might signify the pathology of AD, which, as mentioned earlier, may begin many years or decades before the disease onset (see Section 5.3 below). However, this does not account for the gradual recovery that followed.

As expected, James's syntactic results follow the patterns typical of healthy elders. Her syntactic results vary only slightly, the widest span being an insignificant increase in passive sentence proportion.

Christie's syntactic-complexity results fluctuate in a relatively wider range. The overall trends indicate a rising tendency in all measures, whereas we had predicted declines, although only a few yield significant results. If Christie indeed had AD (as the lexical analysis suggests), our findings are consistent with the observations of Bates *et al.* (1995) that declines in syntax in AD occur in specific, highly constrained areas of complexity, such as passives, and will be observed only in highly constrained situations, such as those that present a natural context for passive sentences. Novels impose no such constraints. Murdoch's and Christie's passive results, however, follow the same direction that Bates *et al.* (1995) documented for AD (a decrease in the passive sentence proportion, a rise in *get*-passives, and a drop in *be*-passives), although 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.

5.3 Murdoch's 'trough'

As noted, an exception to all our expectations is the period of decline in Murdoch's late 40s and early 50s. That dementia-induced language markers in Murdoch's last novel are prefigured by comparable changes in her mid-career suggests that her AD had a long preclinical period. Some circumstantial evidence supports this. On 26 July 1970, Murdoch wrote in her journal: ''I have very little sense of my own identity. Cd one gradually go mad by slowly slowly losing all one's sense of identity? I know there is a body that moves about and some thoughts, memories—but it's all scattered, & now more so'' (Conradi, 2001, pp. 19, 654). A family disposition may have asserted itself in 1975, when

Murdoch's mother became demented (she died ten years later). Murdoch had some long-term episodic memory losses, particularly of the lovers she had in youth (Conradi, 2001, p. 580). In 1993, she abandoned her book on the philosopher Heidegger in page proofs, claiming that it was "no good" (Conradi, 2001, p. 586). Her unofficial biographer observed that she deteriorated as a novelist and philosopher about 1982 because of an "emotional breakdown" and had always exhibited a "general dottiness" (Wilson, 2003, pp. 99, 135, 232). However, we have no explanation for her subsequent partial linguistic recovery prior to her final decline.

5.4 Text analysis for early diagnosis of Alzheimer's disease

Although the scale of our study is much larger than that of previous longitudinal studies of changes in language in AD, we obviously cannot, from just three subjects (one of them of uncertain diagnosis) draw general conclusions about the feasibility of using text analysis to facilitate the diagnosis of AD. Indeed, the wide individual variation, seen even in our small sample, indicates the difficulty of the task. Nonetheless, we found sufficiently clear trends to demonstrate that the idea has merit and is worthy of further development.

5.5 Limitations of our analysis

The analysis that we have made is not without its limitations. Those in data preparation may include typographical errors in the sources or OCR errors, incorrect determination of sentence boundaries caused by missing or incorrect punctuation in the data or by stylistic differences in punctuation usage, and parse trees with either incorrect part-of-speech tags or wrong embedding levels for linguistic components. (However, common patterns of error in the data were identified and corrected.) Errors may also occur in the pattern-matching stage for the D-Level and the passive-voice measures. The pattern sets are not comprehensive, since several kinds of sentence structures cannot be distinguished from others by syntax alone. We excluded highly ambiguous structures from the pattern sets to avoid creating false positives, at the cost of allowing some false negatives. Nonetheless, the defined patterns may also match a relatively small number of false positives.

6 Conclusion

We have presented a large-scale longitudinal study of changes in language in AD using complete, fully parsed texts and a large number of measures, avoiding the limitations and deficiencies of Garrard *et al.*'s (2005) study of Iris Murdoch. We were thereby able to observe trends, such as Murdoch's 'trough', that they could not. Our results support our hypothesis that signs of dementia can be found in diachronic analyses of patients' writings, and in addition lead to new understanding of the work of the individual authors whom we studied. A major loss in vocabulary (revealed by type/token ratio and word-type introduction rate), an increase in repetition of fixed phrases and of content words within close distance, a deficit in noun tokens and a compensation in verb tokens, and a pronounced increase in fillers demonstrate a strong linguistic decline in both Murdoch's and Christie's later works. In contrast, our measures showed no such decline in the language of P.D. James, who is known to have remained healthy. A disease-related linguistic decline is thus clearly distinguished from the effects of healthy aging, as hypothesized. Low-specificity nouns and verbs, however, behave contrary to expectation: a decreasing (non-AD) trend in both Murdoch and James, and an increase in Christie. Also contrary to our hypotheses, both Christie's and James's syntactic complexity results exhibit a slight *rising* tendency (also non-AD), although few measures discover a significant trend. Passivization by Murdoch (a decrease in *be*-passives) and Christie (an increase in *get*-passives) resembles patterns observed in AD patients by Bates et al. (1995), but most results lack statistical significance. However, the syntactic results can be interpreted as evidence that, in AD patients, syntax resists change longer than lexis, as expected.

In future work in this project, we plan first to develop additional measures for comparison, including better measures of word specificity and a repetitiveness index that factors in phrase length, and to look at other aspects of syntactic complexity such as gapping and conjunction, which are reduced in AD (Bates *et al.*, 1995). We will also look at semantic indicators,[5] and in particular, we aim to develop and apply measures of semantic coherence. Also, we are adding new authors to our corpus, including Enid Blyton and Ross Macdonald (pseudonym of Kenneth Millar), both of whom were diagnosed with AD, and additional control authors. To develop this work for clinical use, however, we must demonstrate that the effects of AD can be seen not just in literary text but in ordinary functional daily writing (office memoranda, letters, etc), and it will be necessary to identify current AD patients (and matched controls) with sufficient pre-diagnosis text archives to be used in such a test. For future generations, who are already building a lifetime archive of electronic communications, the availability of past text for eventual diagnosis will not be a problem, though it may be necessary to adapt the measures to the linguistic characteristics of the communication media.

Acknowledgements

We are grateful to Youngchan Kim and Vanessa Wei Feng for pilot studies of this work; to Timothy Harrison for OCR scanning; and to Jackie Chi Kit Cheung, Ming Chow, Elaina Kuang, Ester Macedo, Yaroslav Riabinin, Joshua Sumali, and Amin Tootoonchian for manual error correction of the OCR scans.

Funding

This work was financially supported by a Google Research Award, the Canada Foundation for Innovation, the Social Sciences and Humanities Research Council of Canada, and the Natural Sciences and Engineering Research Council of Canada.

References

Bates, E., Harris, C., Marchman, V., Wulfeck, B., and Kritchevsky, M. (1995). Production of complex syntax in normal ageing and Alzheimer's disease. *Language and Cognitive Processes*, **10**(5): 487–539.

Bird, H., Ralph, M. A. L., Patterson, K., and Hodges, J. R. (2000). The rise and fall of frequency and imageability: noun and verb production in semantic dementia. *Brain and Language*, **73**: 17–49.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly. http://www.nltk.org/book (accessed 1 May 2011).

Blazer, D. G. and Steffens, D. C. (2009). *The American Psychiatric Publishing Textbook of Geriatric Psychiatry*. 4th edn. Arlington, VA: American Psychiatric.

Brown, C., Snodgrass, T., Kemper, S. J., Herman, R., and Covington, M. A. (2008). Automatic measurement of propositional idea density from part-of-speech tagging. *Behavior Research Methods*, **40**(2): 540–5.

Burke, D. M. and Shafto, M. A. (2008). Language and aging. In Craik, F. I. M. and Salthouse, T. A. (eds), *The Handbook of Aging and Cognition*. 3rd edn. New York: Psychology Press, pp. 373–443.

Cantos Gómez, P. (2010). Analyzing the oral speech of an Alzheimer affected person: a case study. *Proceedings, 10th International Conference on Statistical Analysis of Textual Data (JADT 2010)*, Rome, pp. 897–905.

Chan, A. S., Butters, N., and Salmon, D. P. (1997). The deterioration of semantic networks in patients with Alzheimer's disease: a cross-sectional study. *Neuropsychologia*, **35**(3): 241–48.

Charniak, E. (2006). Charniak parser. http://www.cs.brown.edu/~ec/ (accessed 20 March 2011).

Cheung, H. and Kemper, S. (1992). Competing complexity metrics and adults' production of complex sentences. *Applied Psycholinguistics*, **13**: 53–76.

Christie, A. (1977). *Agatha Christie: an Autobiography*. New York: Dodd, Mead.

Conradi, P. J. (2001). *Iris Murdoch: a Life*. New York: Norton.

Cook, C., Fay, S., and Rockwood, K. (2009). Verbal repetition in people with mild-to-moderate Alzheimer disease: a descriptive analysis from the VISTA Clinical Trial. *Alzheimer Disease and Associated Disorders*, **23**(2): 146–51.

Covington, M. A. (2007). CPIDR 3 user manual. Research report 2007-03. CASPR. http://www.ai.uga.edu/caspr/CPIDR-Manual.pdf (accessed 12 August 2010).

Covington, M. A., He, C., Brown, C., Naci, L., and Brown, J. (2006). How complex is that sentence? A proposed revision of the Rosenberg and Abbeduto D-Level Scale. Research report 2006-01. CASPR. http://www.ai.uga.edu/caspr/2006-01-Covington.pdf (accessed 20 March 2011).

Garrard, P., Maloney, L. M., Hodges, J. R., and Patterson, K. (2005). The effects of very early Alzheimer's disease on the characteristics of writing by a renowned author. *Brain*, **128**(2): 250–60.

James, P. D. (2010). P.D. James: The official website. http://www.randomhouse.com/features/pdjames (accessed 2 August 2010).

Kemper, S., Greiner, L. H., Marquis, J. G., Prenovost, K., and Mitzner, T. L. (2001). Language decline across the life span: findings from the Nun Study. *Psychology and Aging*, **16**(2): 227–39.

Lancashire, I. (2010). *Forgetful Muses: Reading the Author in the Text*. Toronto: University of Toronto Press.

Lancashire, I. and Hirst, G. (2009). Vocabulary changes in Agatha Christie's mysteries as an indication of dementia: a case study. *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 (accessed 1 May 2011).

Lancashire, I., Bradley, J., McCarty, W., Stairs, M., and Wooldridge, T. R. (1996). *Using TACT with electronic texts: a guide to Text-Analysis Computing Tools, version 2.1 for MS-DOS and PC DOS*. New York: Modern Language Association of America. http://www.mla.org/store/CID7/PID236 (accessed 1 May 2011).

Le, X. (2010). *Longitudinal Detection of Dementia through Lexical and Syntactic Changes in Writing*, Master's thesis, Department of Computer Science, University of Toronto. http://ftp.cs.toronto.edu/pub/gh/Le-MSc-2010.pdf (accessed 1 May 2011).

Maxim, J. and Bryan, K. (1994). *Language of the Elderly: A Clinical Perspective*. London: Whurr.

Morgan, J. (1984). *Agatha Christie: a Biography*. London: Collins.

Nicholas, M., Obler, L. K., Albert, M. L., and Helm-Estabrooks, N. (1985). Empty speech in Alzheimer's disease and fluent aphasia. *Journal of Speech and Hearing Research*, **28**: 405–10.

Project Gutenberg. http://www.gutenberg.org/ (accessed 1 May 2011).

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

Snowdon, D. A., Kemper, S. J., Mortimer, J. A., Greiner, L. H., Wekstein, D. R., and Markesbery, W. R. (1996). 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): 528–32.

Thomas, C., Keselj, V., Cercone, N., Rockwood, K., and Asp, E. (2005). Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech. *Proceedings of the IEEE International Conference on Mechatronics and Automation,* Niagara Falls, Canada, pp. 1569–74.

Todd, R. (2001). Realism disavowed? Discourses of memory and high incarnations in *Jackson's Dilemma*. *Modern Fiction Studies*, **47**(3): 674–95.

Van Velzen, M. and Garrard, P. (2008). From hindsight to insight — retrospective analysis of language written by a renowned Alzheimer's patient. *Interdisciplinary Science Reviews*, **33**(4): 278–86.

Williams, K., Holmes, F., Kemper, S., and Marquis, J. (2003). 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): 42–4.

Wilson, A. N. (2003). *Iris Murdoch as I knew her*. London: Hutchinson.

Yngve, V. H. (1960). A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society*, **104**(5): 444–66.

Notes

1 Snowdon *et al.* (1996) found a strong and consistent association between low idea density in texts written early in life 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.

2 Nonetheless, Todd (2001), while acknowledging that the novel possibly contains lapses, defends its literary qualities.

3 A subsequent study by van Velzen and Garrard (2008) of three novels by the author Gerard Reve, also known to have died of Alzheimer's disease, looked only at changes in the author's lexical diversity, and concluded that the results for this measure were similar to those found for Murdoch. Similarly, Cantos (2010) found an increase over time in *n*-gram repetitions in the Parliamentary speeches of former British Prime Minister Harold Wilson, whose eventual Alzheimer's disease is believed to have shown symptoms prior to his resignation. Thomas *et al.* (2005) achieved moderate accuracy using authorship analysis techniques to discriminate severity of dementia in AD patient interview transcripts.

4 Christie also wrote romance novels under the name Mary Westmacott.

5 Although reduction in the idea density of discourse has been associated with aging and AD (Burke and Shafto, 2008), a pilot study by our colleague Vanessa Wei Feng using CPIDR 3 (Covington, 2007; Brown *et al.*, 2008) showed no significant change over time of idea density in the writing of Christie, James, or Murdoch.

Appendix

List of Novels

Iris Murdoch

Novel	Approx. Age at Composition	Year of Publication
Under the Net	35	1954
The Flight from the Enchanter	36	1955
The Bell	39	1958
A Severed Head	42	1961
An Unofficial Rose	43	1962
The Unicorn	44	1963
The Italian Girl	45	1964
The Time of the Angels	47	1966
The Nice and the Good	49	1968
Bruno's Dream	50	1969
A Fairly Honorable Defeat	51	1970
The Black Prince	54	1973
The Sacred and Profane Love Machine	55	1974
Henry and Cato	57	1976
The Sea, the Sea	59	1978
The Philosopher's Pupil	64	1983
The Good Apprentice	66	1985
The Book and the Brotherhood	68	1987
The Green Knight	74	1993
Jackson's Dilemma	76	1995

Agatha Christie

Novel	Approx. Age at Composition	Year of Publication	Technology
The Mysterious Affair at Styles	28	1920	T
The Secret Adversary	32	1922	T
The Murder of Roger Ackroyd	34	1926	T
Murder on the Orient Express	43	1934	T
Appointment with Death	47	1937	T
Curtain*	50	1975	T
Towards Zero	51	1944	T
A Murder is Announced	59	1950	T
Destination Unknown	63	1954	D
Ordeal by Innocence	67	1958	D
The Clocks	72	1963	D
Endless Night	76	1967	D
Passenger to Frankfurt	79	1970	D
Nemesis	80	1971	D
Elephants Can Remember	81	1972	D
Postern of Fate	82	1973	DE

T: Typewriter; D: Dictaphone; DE: Dictaphone + Editing
*Written between 1940 and 1941.

P.D. James

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