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

Le et al. (2011)

Daniela Stier

University of Tübingen

daniela.stier@student.uni-tuebingen.de

December 2, 2015

Table of Contents

Motivation

- Alzheimer's disease (AD) among most prevalent geriatric conditions
- definite diagnosis only *post mortem*
- no proven cure for dementia

→ correct, timely diagnosis is of great importance

→ sufficiently early diagnosis of AD may even make prevention possible in future

Background

> *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.*
>
> *(Blazer et al., 2004, p. 249)*

- early diagnosis through **linguistic analysis**:
  - affecting linguistic abilities in speech and writing
  - possibility to develop techniques, e.g. looking for diachronic changes in patients' writings
  - problem: how to get a lifelong corpus of writing?

Approach

- **study:** large-scale longitudinal study of lexical and syntactic changes in language in Alzheimer's disease (AD)
- **corpus:** complete, fully parsed texts by three authors
- **hypothesis:**
  - $\longrightarrow$ signs of dementia can be found in diachronic analyses of patients' writings
  - $\longrightarrow$ lead to a new understanding of the work of the individual authors
- **related:** Williams et al. (2003), Garrard et al. (2005)[1]

[1] Please refer to Le (2010) for more details.

Material

P. D. James
(*1920 - †2014)[2]

aged healthily

Agatha Christie
(*1890 - †1976)[3]

suspected of Alzheimer's

Iris Murdoch
(*1919 - †1999)[4]

died with Alzheimer's

[2] Extracted from http://www.independent.co.uk/news/people/pd-james-dead-detective-novelist-behind-death-comes-to-pemberley-and-children-of-men-dies-aged-94-9887573.html (last access: 11/30/2015).

[3] Extracted from http://www.niederdeutschebuehne.de/agatha-christie/ (last access: 11/30/2015).

[4] Extracted from http://www.theguardian.com/commentisfree/2009/jun/26/iris-murdoch (last access: 11/30/2015).

Material

P. D. James

Agatha Christie

Iris Murdoch

Expectation:

James data will exhibit linguistic patterns of healthy ageing

Christie data will exhibit linguistic patterns similar to those of AD patients

Murdoch data will exhibit linguistic patterns of dementia patients

Assumption:

- no novel 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

Language in Ageing and Dementia

- **consensus:** decline occurring in normal ageing is accelerated in presence of AD
- **AD:** deficits in lexical features may be more prominent than in syntactic ones

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

Figure: Expected patterns of linguistic changes

Lexical Analysis

Variety of measures for **lexical markers**:

a Vocabulary size
b Lexical repetition
c Word specificity
d Word-class deficits
e Fillers

Text length

measures sensitive to length:
→ threshold: 55.000 tokens

for remaining measures:
complete text of all novels

Changes over time: simple linear regression of the respective measure against the author's age

Statistical significance: relationship between the author's age and the value of the respective measure

Spearman correlation: correlation between measures **!**

Lexical Analysis

a) Vocabulary size

- TTR: type/token ratio
- number of unique lemmatized word-types divided by total number of word-tokens

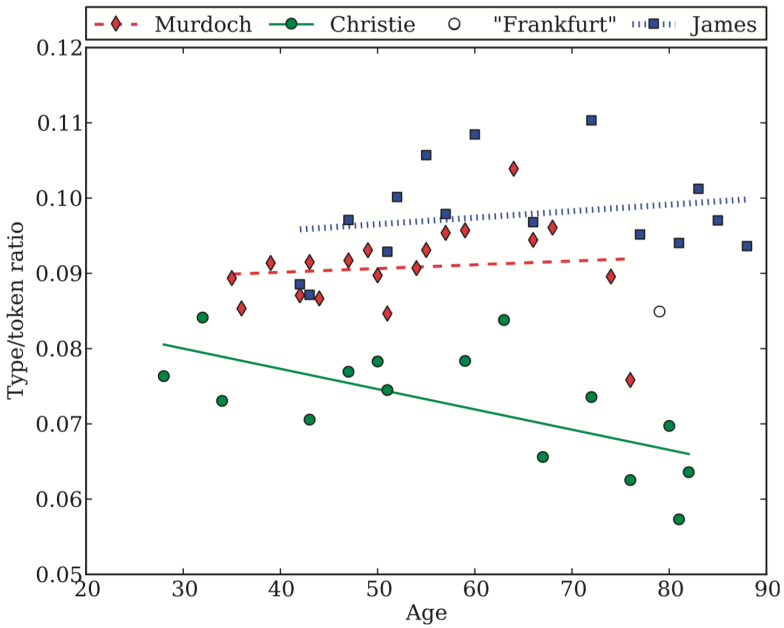


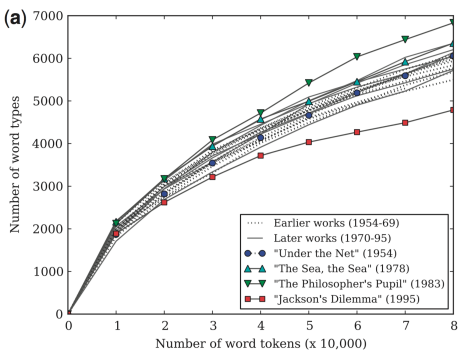
Figure: Type/token ratio within the first 55.000 tokens

- M: rates drop - rise - drop, insignificant
- C: rates vary, drop, significant
- J: slight rising trend, insignificant

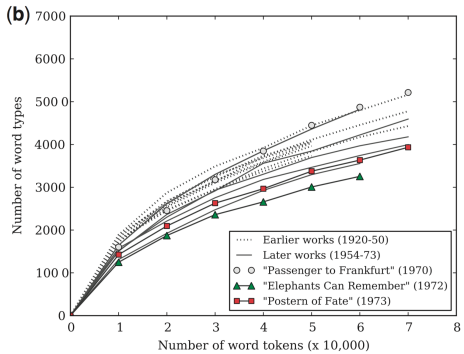
Lexical Analysis

a) Vocabulary size

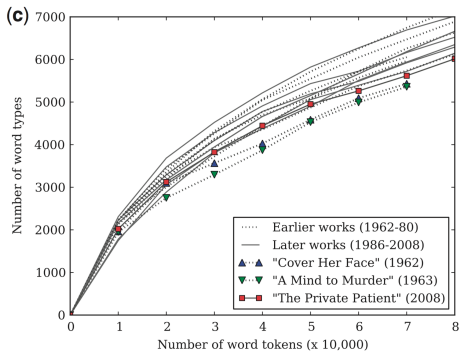
- WTIR: word-type introduction rate
- cumulative number of unique lemmatized types, computed at 10.000-token interval



(a) I. Murdoch



(b) A. Christie



(c) P. D. James

Figure: Word-type introduction rate up to the 70.000th token

Lexical Analysis

b) Lexical repetition

- global word n-gram repetitions
- i.e. 2-11 words, occurring at least twice
- **maximals**: longest repeating phrases in a text
- **associates**: substrings of maximals occurring more frequently than maximals

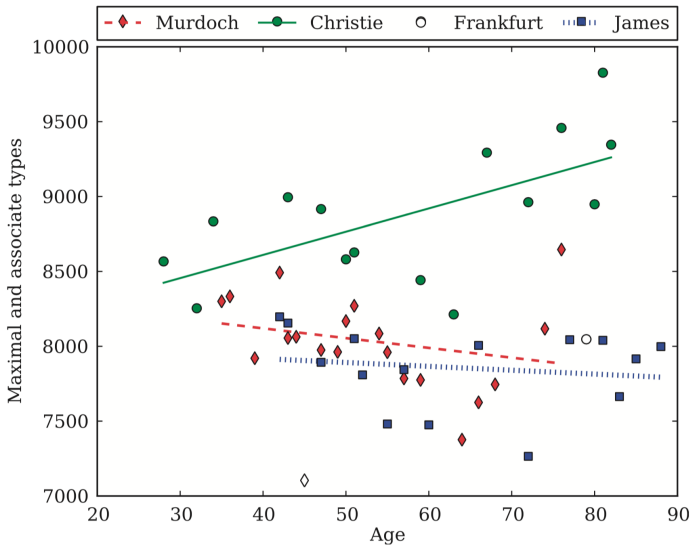


Figure: Maximal and associate phrasal repetitions (types)

- M: rise and peak in last novels, overall decrease, insignificant
- C: overall increase, significant
- J: overall decrease, insignificant

Lexical Analysis

b) Lexical repetition

- local proportion of lemmatized open-class words
- i.e. nouns, content verbs, adjectives, adverbs
- repeated within 10 subsequent open-class words
- computed over the number of all content words in each novel

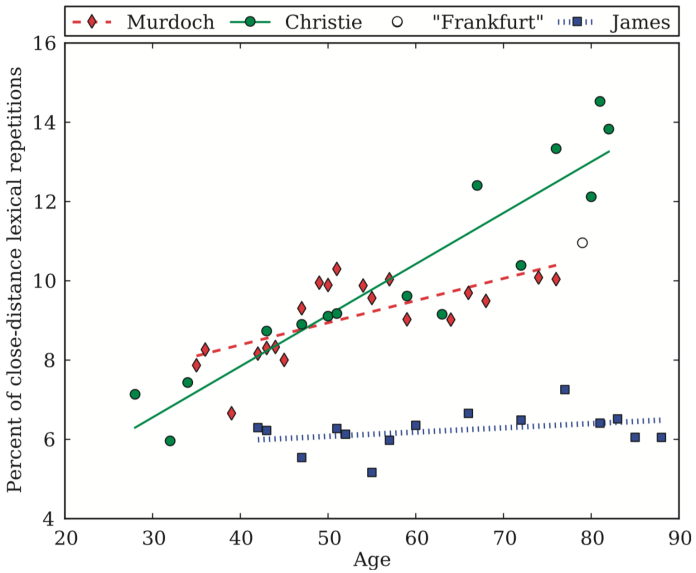


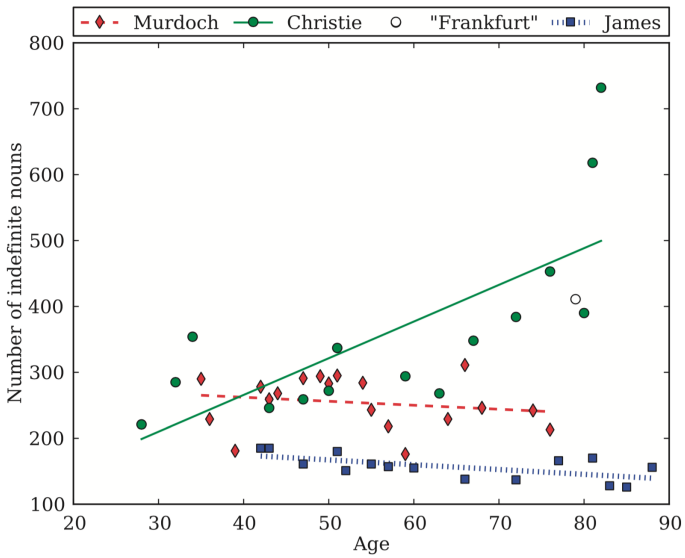
Figure: Lexical repetitions within 10 subsequent content words

- M: overall increase, significant
- C: steep rise, sharp contrast
- J: rate relatively stable

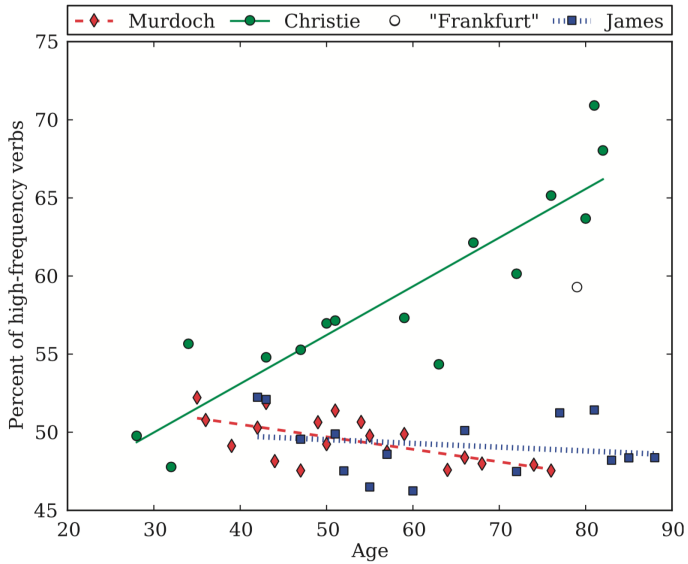
Lexical Analysis

c) Word specificity

- proportions of indefinite nouns and high-frequency, low-imageability verb tokens
- 4 indefinite nouns (*thing*(*s*), *something*, *anything*, *nothing*)
- 35 high-frequency verbs of relatively low specificity (*be*, *come*, *do*, *get*, ...)



(a) Number of indefinite noun occurrences

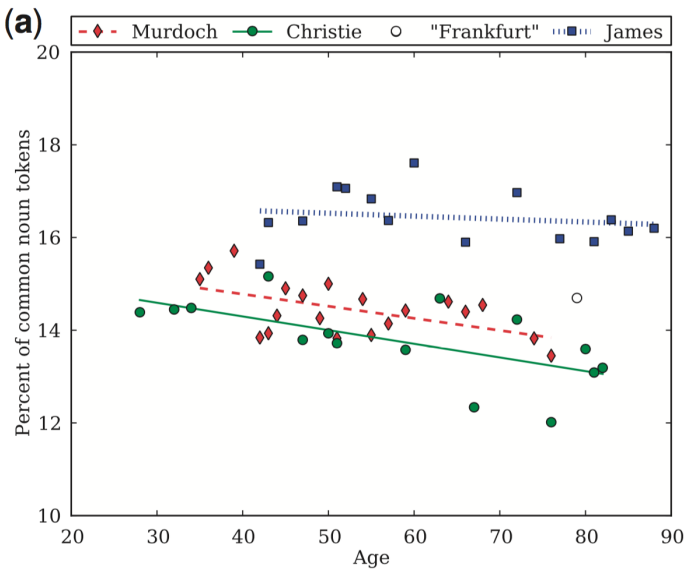


(b) Proportion of 35 high-frequency verbs

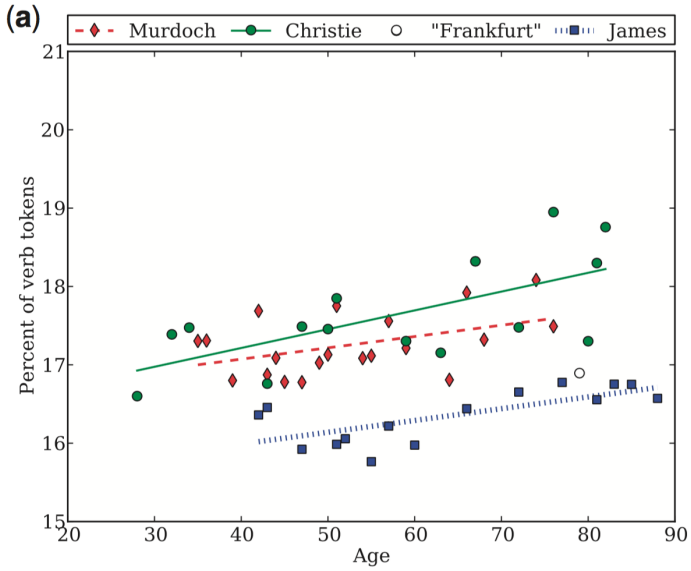
Lexical Analysis

d) Word-class deficits

- proportions of each word class over the entire length of a text
- word-tokens
- word-types



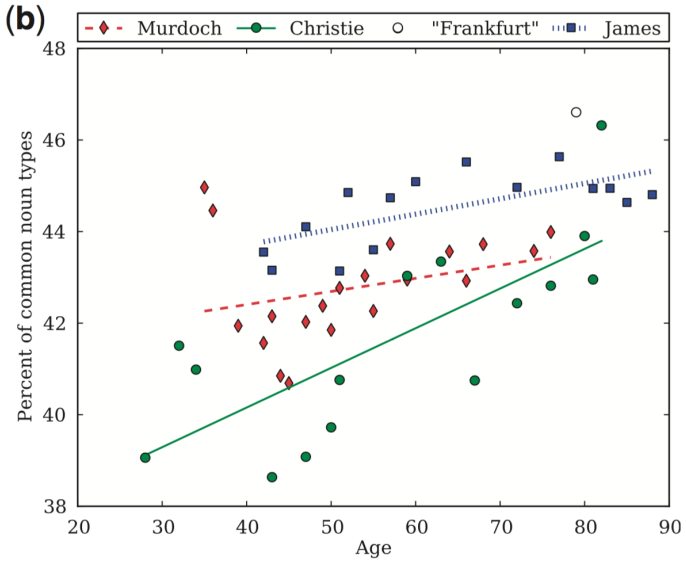
(c) Proportion of common nouns by token



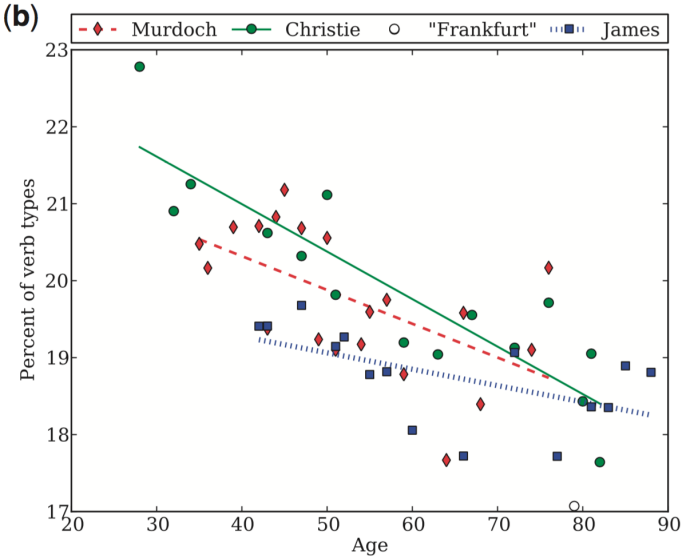
(d) Proportion of content verbs by token

Lexical Analysis

d) Word-class deficits



(e) Proportion of common nouns by type



(f) Proportion of content verbs by type

Lexical Analysis

e) Fillers

- proportion of words, part-of-speech-tagged as interjections and fillers
- caution: might reflect stylistic choices rather than cognitive decline

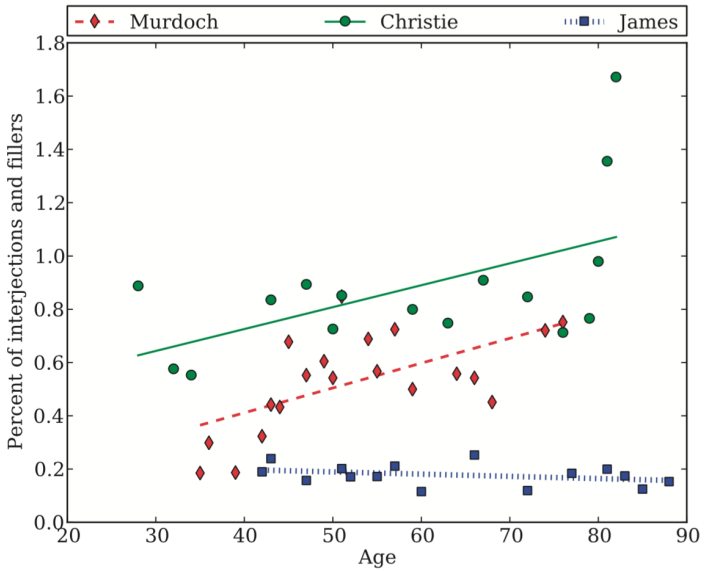


Figure: Proportion of interjection and fillers

- M: rising trend, significant
- C: rising trend, significant
- J: slight decreasing trend, insignificant

Syntactic Analysis

Measures for **syntactic markers**:

a Syntactic complexity
- MLU & MCU
- Parse tree depth
- D-Level

b Passive voice

Parse trees

most measures operate on parse trees

Changes over time: simple linear regression of the respective measure against the author's age

Statistical significance: relationship between the author's age and the value of the respective measure

Spearman correlation: correlation between measures **!**

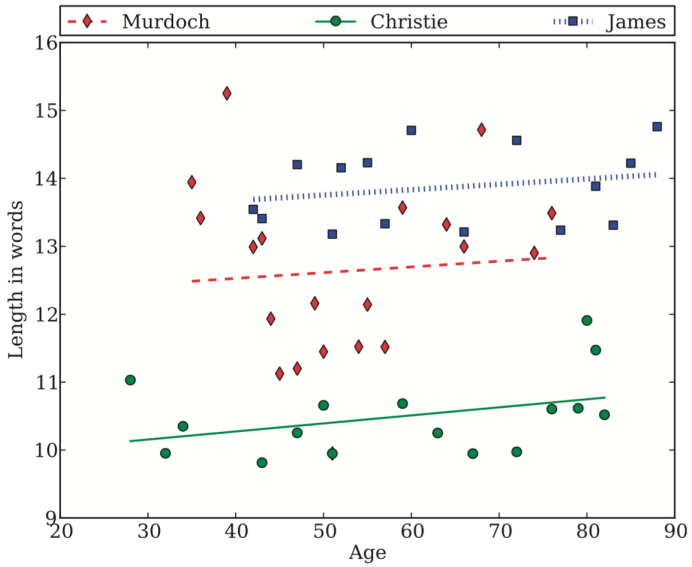
Syntactic Analysis

a) Syntactic complexity: MLU & MCU

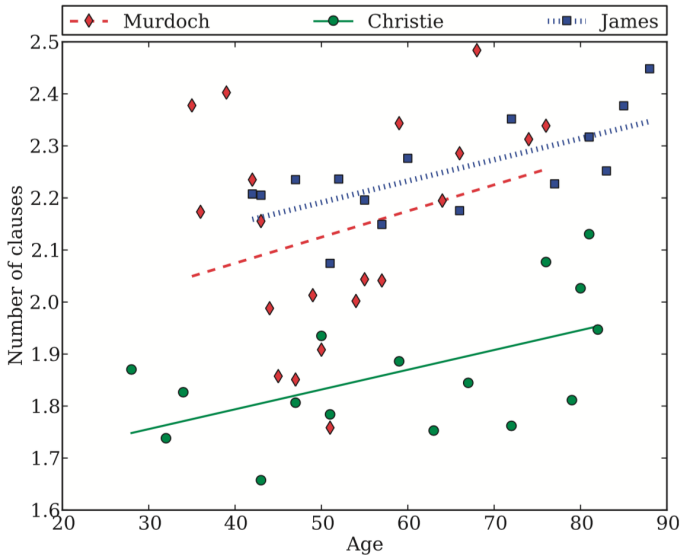
- number of words and number of clauses (main, subordinate, embedded)
- contractions count as two words (e.g. *is* - *n't*)

MLU mean length of utterance

MCU mean number of clauses per utterance



(a) Mean length in words per sentence



(b) Mean number of clauses per sentence

Syntactic Analysis

a) Syntactic complexity: Parse tree depth

- average maximum depths of parse trees
- reflects average number of embedded structures in a sentence

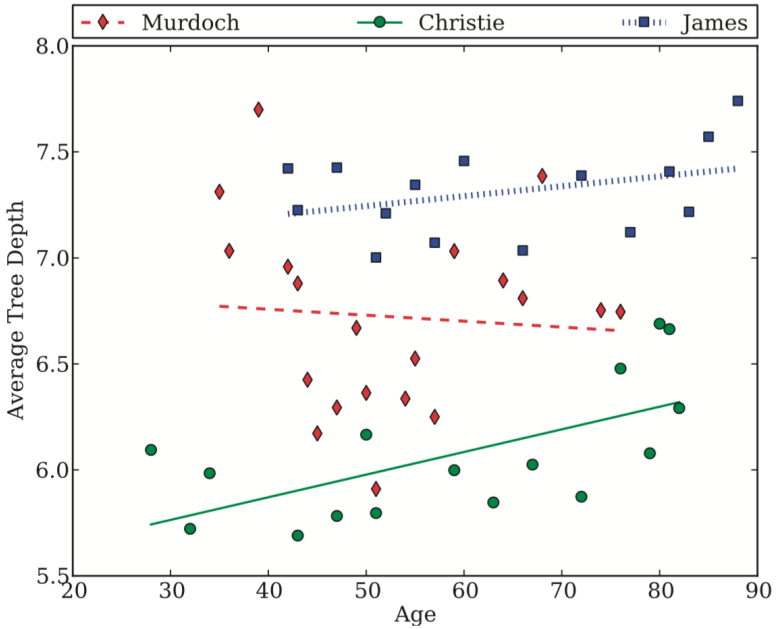


Figure: Average parse tree depth

- M: rates rise - drop - rise, overall decrease, insignificant
- C: overall increase, significant
- J: rates consistent, overall increase, insignificant

Syntactic Analysis

a) Syntactic complexity: D-Level

- psycholinguistics-based ranking of sentences
- 8 levels of increasing syntactic complexity
- pattern-matching to determine the level of a parse tree
- Covington et al. (2006)

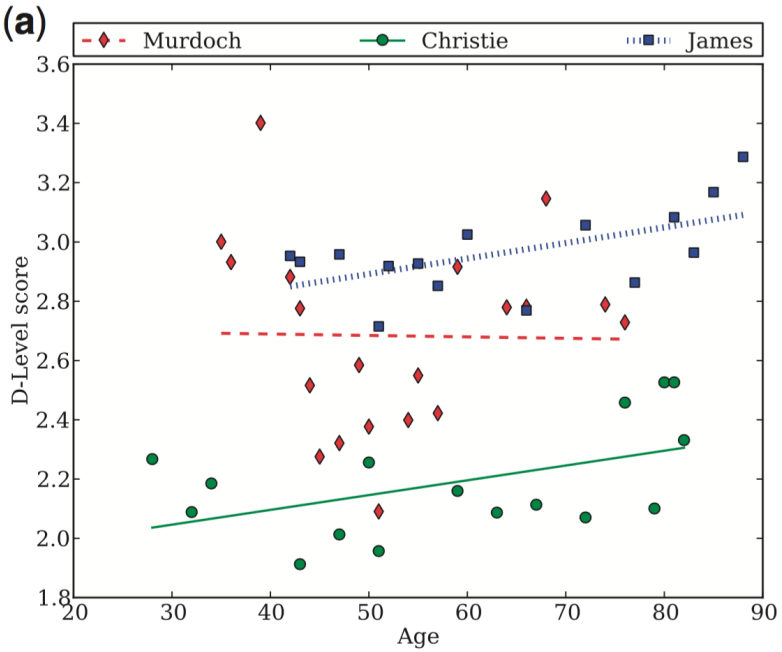


Figure: Average D-Level score

- M: slight decrease, significant
- C: overall increase, significant
- J: overall increase, significant

Syntactic Analysis

b) Passive voice

- number of sentences containing...
  - *be*-passive
  - *get*-passive
  - past participle followed by a *by*-phrase

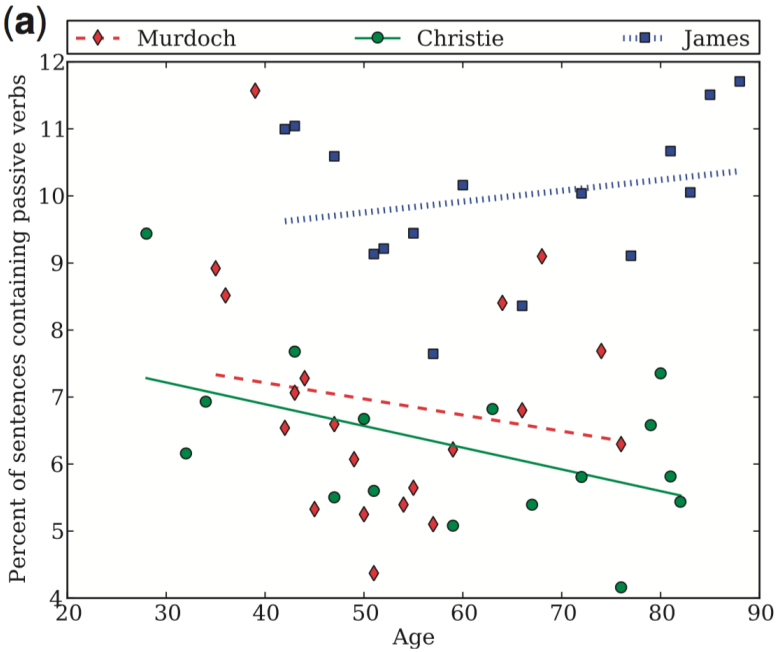


Figure: Proportion of passive sentences

- M: overall decline
- C: decline, significant
- J: slight increase

Summary

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.

Figure: Observed patterns of linguistic changes

Conclusion

- ✓ signs of dementia can be found in diachronic analyses
- linguistic decline in later works of M./C.:
  - loss in vocabulary (TTR, WTIR)
  - increase in repetition of phrases
  - increase of content words within close distance
  - deficit in noun tokens
  - compensation in verb tokens
  - increase in fillers
- no such decline in J's language
- low-specificity nouns and verbs contrary to expectation (decrease = non-AD for M./J., increase for C.)
- syntactic complexity contrary to expectation (rising = non-AD for C./J.)
- syntactic results: in AD syntax resists change longer

$\longrightarrow$ disease-related linguistic decline clearly distinguished from healthy ageing

Conclusion

- limitations: data preparation, pattern-matching (D-Level, passive voice), ...
- three subjects not sufficient to draw general conclusions
- sufficiently clear trends found to demonstrate that further development is useful
- future work: additional measures for comparison, including semantics, etc.

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.

Blazer, D. G., Steffens, D. C., and Busse, E. W. (2004). *The American psychiatric publishing textbook of geriatric psychiatry*. American Psychiatric Pub.

Charniak, E. (2000). A maximum-entropy-inspired parser. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 132–139. Association for Computational Linguistics.

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.

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):250–260.

Le, X. (2010). Longitudinal detection of dementia through lexical and syntactic changes in writing. *Science*.

Le, X., Lancashire, I., Hirst, G., and Jokel, R. (2011). Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three british novelists. *Literary and Linguistic Computing*, page fqr013.

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*, 58(1):P42–P44.

Summary - Lexical Analysis

- results largely support the hypothesis
- **James**
  - results follow patterns expected for normal ageing
  - vocabulary, repetition, specificity only vary in small range
  - no word-class deficit observed
- **Christie**
  - overall decline for vocabulary, repetition, specificity
  - deficit in noun tokens → increase in verb tokens
- **Murdoch**
  - TTR and WTIR in later novels show lexical decline
  - drop in vocabulary size → increase in repetitions of content words
  - deficit in noun tokens → increase in verb tokens
  - repeating phrases rise steadily after her 60s
  - lexical specificity remained intact

Summary - Syntactic Analysis

- **James**
  - results follow patterns expected for normal ageing
  - results only vary slightly
  - widest span an insignificant increase in passive sentence proportion
- **Christie**
  - results fluctuate in a relatively wide range
  - overall rising tendency for all measures → whereas decline was expected!
  - if lexical analysis suggests AD, results in syntax coincide with Bates et al. (1995)[5]
- **Murdoch**
  - no significant linear trends observed
  - all measures reveal abrupt drop in her late 40s/50s, followed by a period of recovery, for some measures followed by a slight decline in last two novels

[5] Declines in syntax in AD occur in highly complex areas, s.a. passives, and will only be observed in highly constrained situations, s.a. natural context for passive sentences.

Background information: Data

P. D. James
(*1920 - †2014)

Agatha Christie
(*1890 - †1976)

Iris Murdoch
(*1919 - †1999)

James
- 15 novels analysed
- published between ages 42 and 82 → M = 63.9

Christie
- 16 novels analysed
- published between ages 28 and 82 → M = 59.9

Murdoch
- 20 novels analysed
- published between ages 35 and 76 → M = 52.7

- each novel assumed to be written just before publication
- all texts belong to the same genre (i.e. *prose fiction*)
- novels span the author's career
- analysed texts without any influence (e.g. by collaborating writer, editor, etc.)

Background information: Data Preprocessing

- novels scanned and converted to plain text (OCR)
- OCR errors in spelling and pronunciation corrected manually

Levels of Text Processing

1. separation of punctuation marks and clitics from word tokens (e.g. *I - 'm*, *John - 's*)
2. lemmatization (NLTK WordNet's morphy method)
3. determination of sentence boundaries (rule-based deterministic algorithm)
4. generation of a parse tree for each sentence (Charniak (2000) parser, includes part-of-speech tagging)
5. correction of common error patterns made by the parser (interactive script)

Background information: Spearman's correlation

Spearman rank-order correlation coefficient:

- statistical dependence betw. two variables
- monotonic function describes relationship betw. two variables

IQ, X_i	Hours of TV per week, Y_i	rank x_i	rank y_i	d_i	d_i^2
86	0	1	1	0	0
97	20	2	6	–4	16
99	28	3	8	–5	25
100	27	4	7	–3	9
101	50	5	10	–5	25
103	29	6	9	–3	9
106	7	7	3	4	16
110	17	8	5	3	9
112	6	9	2	7	49
113	12	10	4	6	36

Figure: Example correlation[6]

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}, \quad (1)$$

$$d_i = x_i - y_i$$

$$\rho = 1 - \frac{6 \cdot 194}{10(10^2 - 1)} \quad (2)$$

$$\Longrightarrow \rho = -0.17575...$$

Figure: Formula and calculation

[6] Extracted from https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient (last access: 12/01/2015).