Natural language processing methods for the detection of symptoms of Alzheimer's disease in writing

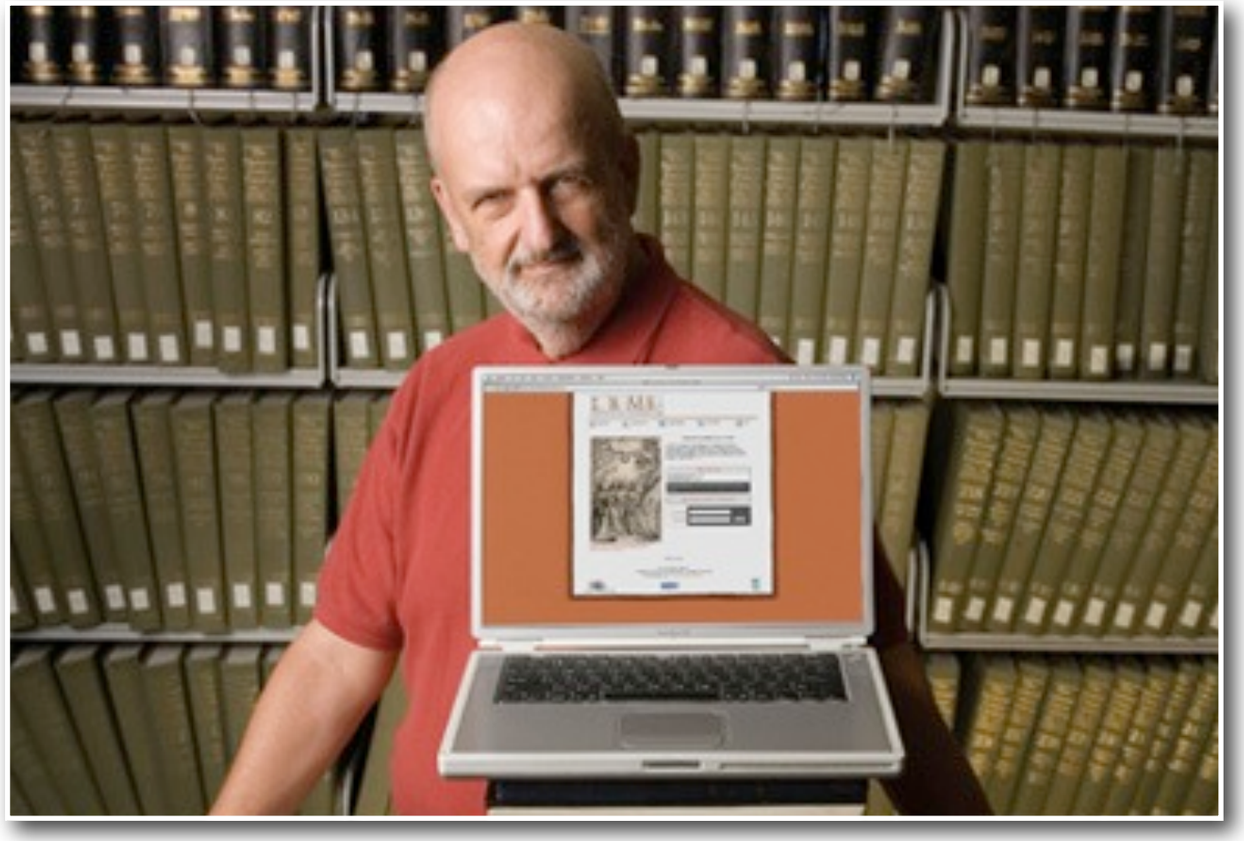
Graeme Hirst
Computer Science
University of Toronto

and ...

Xuan Le

Computer Science

Ian Lancashire *English*



Regina Jokel

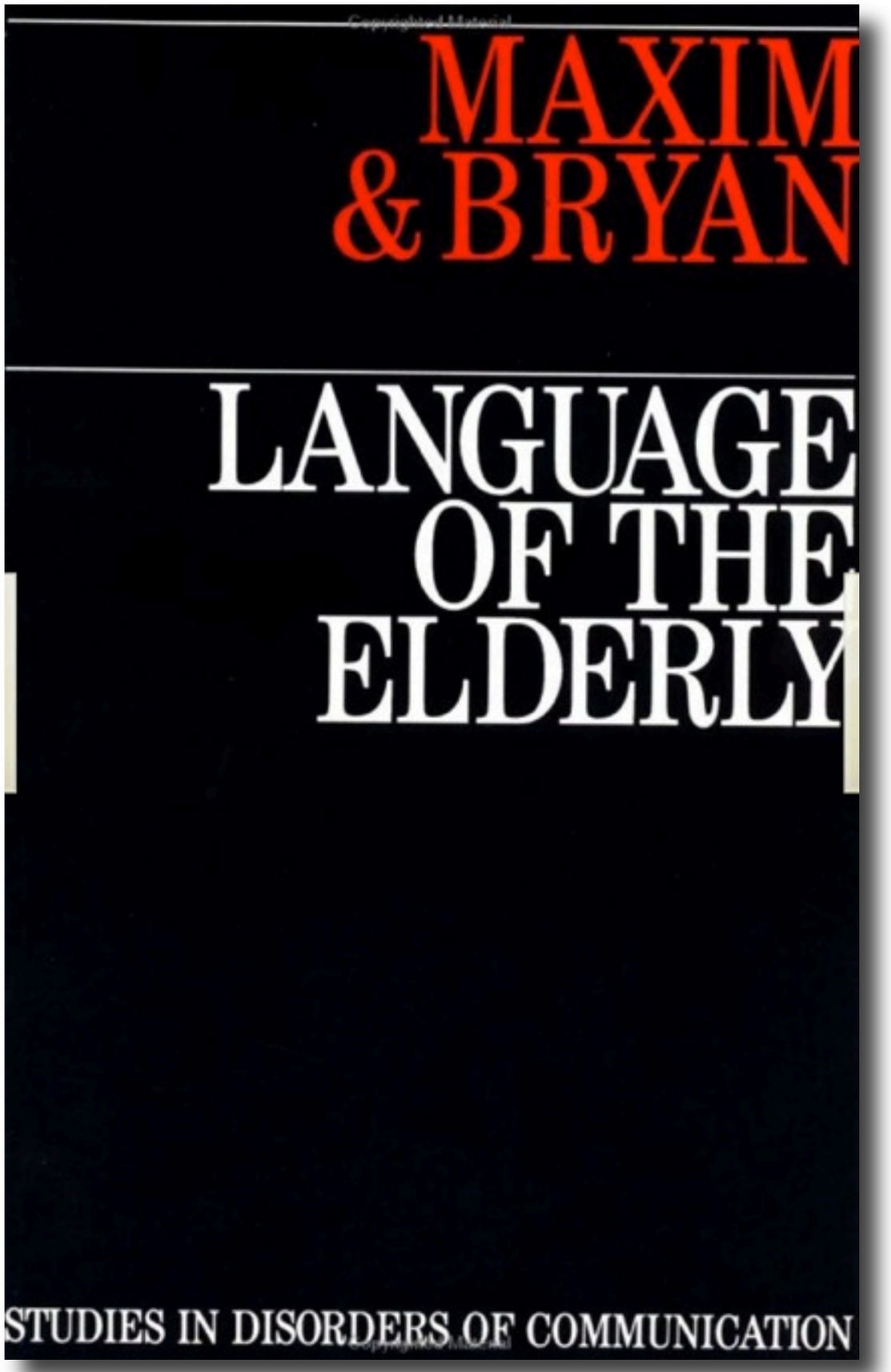
Speech–Language Pathology and
Baycrest Centre for Geriatric Care

With the support of a
University Research Award from

Google

Sponsor: Dekang Lin

1
Language in healthy aging


MAXIM
& BRYAN
LANGUAGE
OF THE
ELDERLY
STUDIES IN DISORDERS OF COMMUNICATION

Language and cognition change naturally throughout life

– Change ≠ deterioration

- Vocabulary expands throughout life.
- Elderly make better use of context and semantic strategies.
- Elderly are better at integrating information.
- Elderly tend to use more-abstract vocabulary.

Other age-related changes affect language use and communication

- Hearing loss
- Reduced visual acuity
- Reduced accuracy of articulation
- Changes in memory and retrieval processes

Changes in language production and comprehension strategies

Production by elderly

- More performance errors, more disfluency.
  More fillers (due to retrieval problems?).
  Greater use of indefinite words.
  Slower at producing names in a category.
  Greater tip-of-tongue for common nouns but not abstract ones.
  Less-complex syntax, shorter sentences.
- No changes in discourse competence.
  Better in synonyms test.
- But overall, differences are small.

- Maxim, Jane and Bryan, Karen. *Language of the Elderly: A clinical perspective*. Whurr, 1994.

Understanding by elderly

- Use frequency of sentence type more.
  Use order of mention more.
  Use semantics more.
  Prefer main-clause-first sentences.
  Make more use of context.
  Have less memory for modifiers and logical connectives.
  Find it harder to spot anomalies.

- But overall, differences are small.

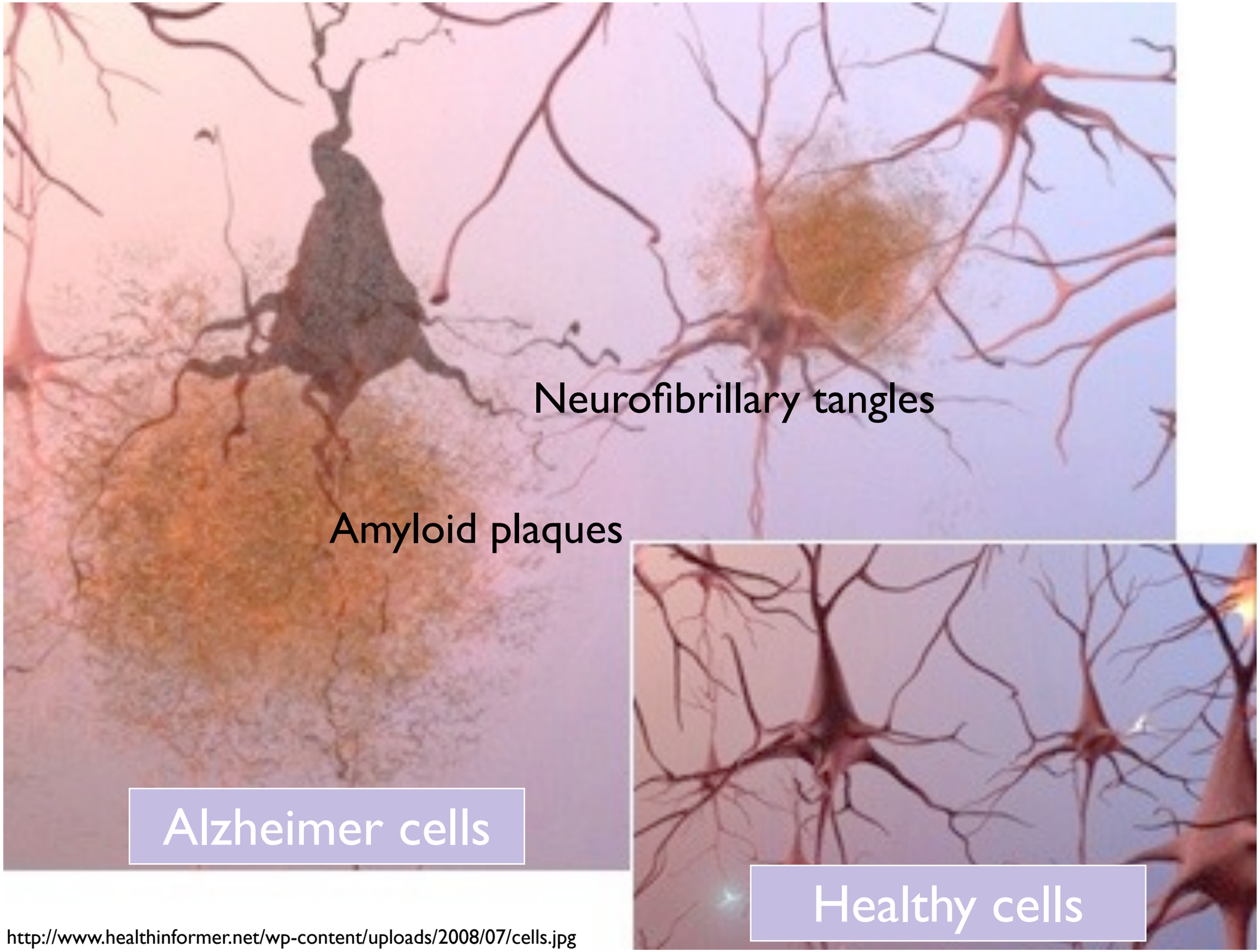
- Maxim, Jane and Bryan, Karen. *Language of the Elderly: A clinical perspective*. Whurr, 1994.

2
Language in Alzheimer's disease

Alzheimer's dementia

- Most common form of dementia.
- Caused by Alzheimer's disease.
  - Cortical degeneration.

- Lyketsos, Constantine G. Dementia and milder cognitive syndromes. In: Blazer, Dan G. and Steffens, David C. (eds). *The American Psychiatric Publishing Textbook of Geriatric Psychiatry,* 4th ed., 2009.
- Maxim, Jane and Bryan, Karen. *Language of the Elderly: A clinical perspective.* Whurr, 1994.



http://www.healthinformer.net/wp-content/uploads/2008/07/cells.jpg

Alzheimer's dementia

- Most common form of dementia.
- Caused by Alzheimer's disease.
  - Cortical degeneration.

- Progressive irreversible decline in many areas of cognition.
  - Language comprehension and production, memory, problem-solving, social skills, …

- Lyketsos, Constantine G. Dementia and milder cognitive syndromes. In: Blazer, Dan G. and Steffens, David C. (eds). *The American Psychiatric Publishing Textbook of Geriatric Psychiatry,* 4th ed., 2009.
- Maxim, Jane and Bryan, Karen. *Language of the Elderly: A clinical perspective.* Whurr, 1994.

Alzheimer's dementia

- Initial stage: mild cognitive impairment (MCI).
  - 80% of cases: eventual AD.
- Cognitive assessment: MMSE, 3MS, etc.
- Can be hard to differentiate from some other dementias prior to post-mortem.
  - Often, multiple pathologies are present.
  - Wide individual variation in symptoms.

- Lyketsos, Constantine G. Dementia and milder cognitive syndromes. In: Blazer, Dan G. and Steffens, David C. (eds). *The American Psychiatric Publishing Textbook of Geriatric Psychiatry,* 4th ed., 2009.

Alzheimer's dementia

- No cure, but progression can be slowed, perhaps even halted, by medication.

  "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." Lyketsos 2009

  Or linguistic markers?!

- Lyketsos, Constantine G. Dementia and milder cognitive syndromes. In: Blazer, Dan G. and Steffens, David C. (eds). *The American Psychiatric Publishing Textbook of Geriatric Psychiatry,* 4th ed., 2009.

Lexical changes

Marker	Dementia	Healthy aging
Vocabulary size	Sharp decrease	Gradual increase, then possible slight decrease
Lexical repetition	Pronounced increase	Possible small change
Word specificity	Pronounced decrease	Possible small change
Word class distribution	Fewer nouns, compensation in verbs	No change
Fillers	Pronounced increase	Possible slight increase

Syntactic changes

Marker	Dementia	Healthy aging
Syntactic complexity	Sharp decline	Little or no change, then possible rapid decline in mid-70s
Use of passive voice	Pronounced decrease	Possible small decrease
Auxiliary verb in passive voice	*Get* dominates *be*	*Be* dominates *get*
Passives without agent	Greater decrease	Moderate decrease

Idea:

Your word processor could watch out for changes in your writing.

Premises:

– Linguistic changes marking dementia can be reliably seen in a writer's text.

– User has a lifetime of text for comparison with new writing.

Problem:

Testing this idea in 2010.

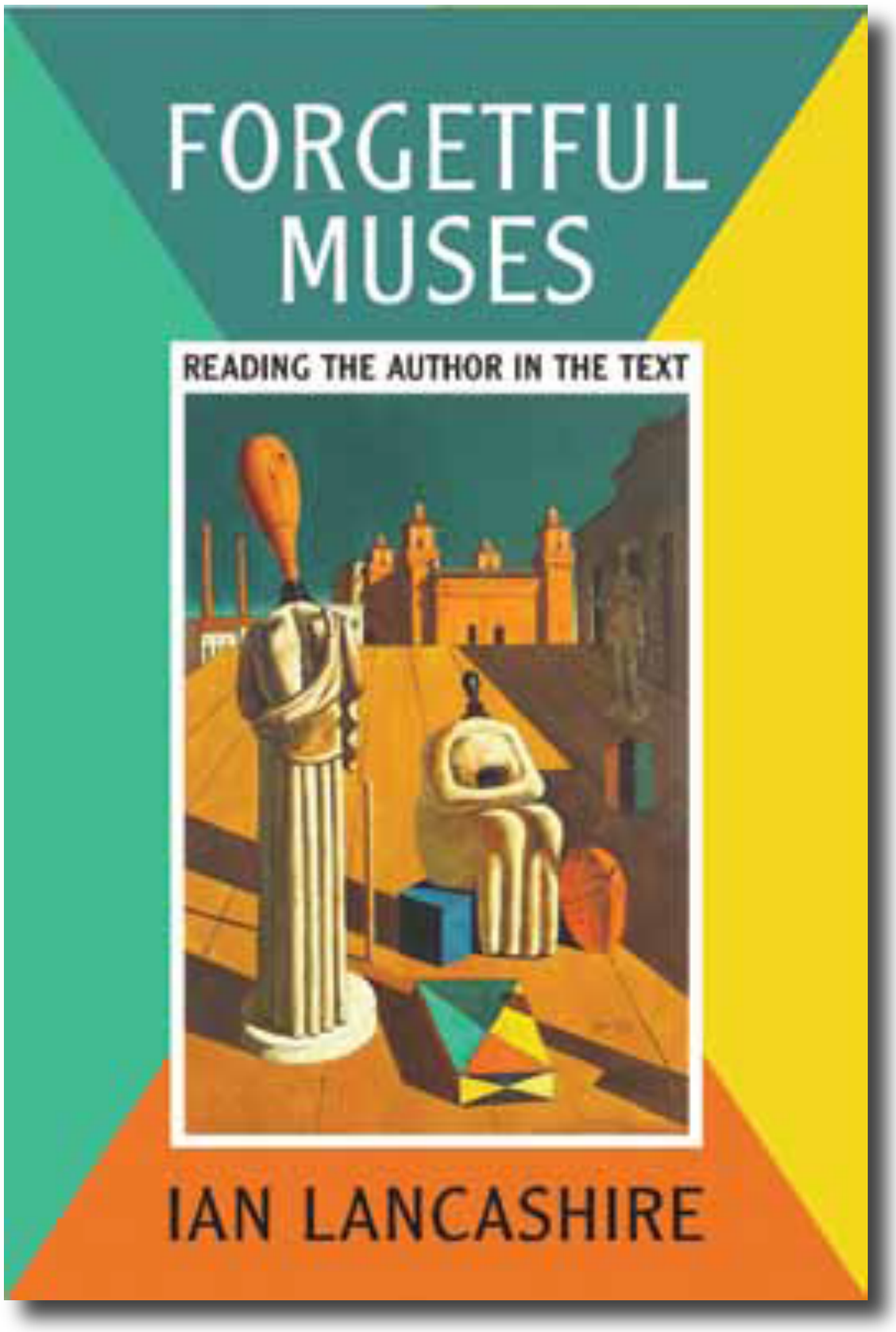
Requirements:

Subjects with lifetime text corpus and confirmed Alzheimer's status.

Solution:

Writers with large published œuvre.

3
The author in the text



To be published
October 2010

Death of death-of-the-author

- Quantitative and qualitative textual analysis to find author's unconscious traits.
  - *cf.* quantitative methods in authorship attribution by stylistic markers.
  - Emphasis now on discovering individual author's cognitive processes of writing and intention.

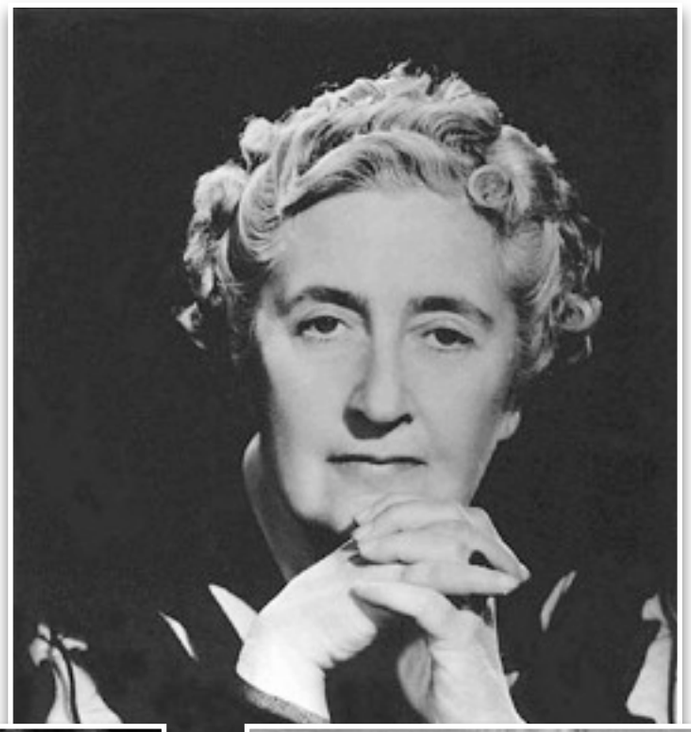
- Wimsatt, William K. Jr. and Beardsley, Monroe C. 1954. "The intentional fallacy." *The Verbal Icon: Studies in the Meaning of Poetry*. Lexington: University of Kentucky Press. 3–18.
- Barthes, Roland. "The death of the author." *Aspen*, 5+6, Fall–Winter 1967, item 3.
- Fish, Stanley. *Is There a Text in this Class? The Authority of Interpretive Communities*. Harvard University Press, 1980.

A special case of the author in the text

The author whose cognitive processes are damaged

4
Data and hypotheses

Agatha Christie, 1890–1976



Agatha Christie, 1890–1976

- British detective novelist.
- Intricate plots.
- 90 novels over 53 years.
- Third best-selling author of all time
  (after Bible and Shakespeare).
- Final novels poor in quality, dull, contain errors.
- Suspected cognitive decline.

Iris Murdoch, 1919–1999

Iris Murdoch, 1919–1999

- Acclaimed British novelist and philosopher.
- 26 novels over 41 years.
- Final novel was a complete mess.
- Diagnosed with Alzheimer's disease.
- Minimal amateur linguistic analysis by Garrard et al (2005).

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

P.D. James, 1920–

January 2010

P.D. James, 1920–

- British detective novelist (and SF).
- Literary writer.
- 21 novels over 48 years.
- Still publishing and fully active in her late 80s. No evidence of cognitive decline.

Hypotheses:

Murdoch and Christie will show changes as in dementia.

James will show little or no change, as in normal aging.

In all cases, we are looking for **relative change** within an individual, not at absolute numbers.

Data

Author	No of novels	Age at composition
Christie	16	28 to 82
Murdoch	20	35 to 76
James	15	42 to 88

Data

- Editing by publisher?
  - Christie and Murdoch: Little or none.
    (except Christie's last novel was 'cleaned up' somewhat by family friends).
  - James: Unknown, probably little.

Data

- Effects of writing process, genre, and style?
  - Christie: Typewriter versus dictaphone.
  - Christie: *Passenger to Frankfurt* — Late career spy novel based on large amounts of reading.
    Processed but excluded from most analyses.
    Pseudonymous romances by 'Mary Westmacott' excluded.
  - James: Science fiction novel *The Children of Men*.
  - Murdoch: Variety of topics, sometimes told in voice of central character.
    Short novel *The Italian Girl* excluded from length-dependent analyses.

Preparation

- Data not legally available from Google Books.
- Canadian copyright law permits unrestricted copying for research.
  Fair-use defense in U.S. law probably not adequate here.
- OCR, human error correction
  (except two Christies from Project Gutenberg).

Preparation

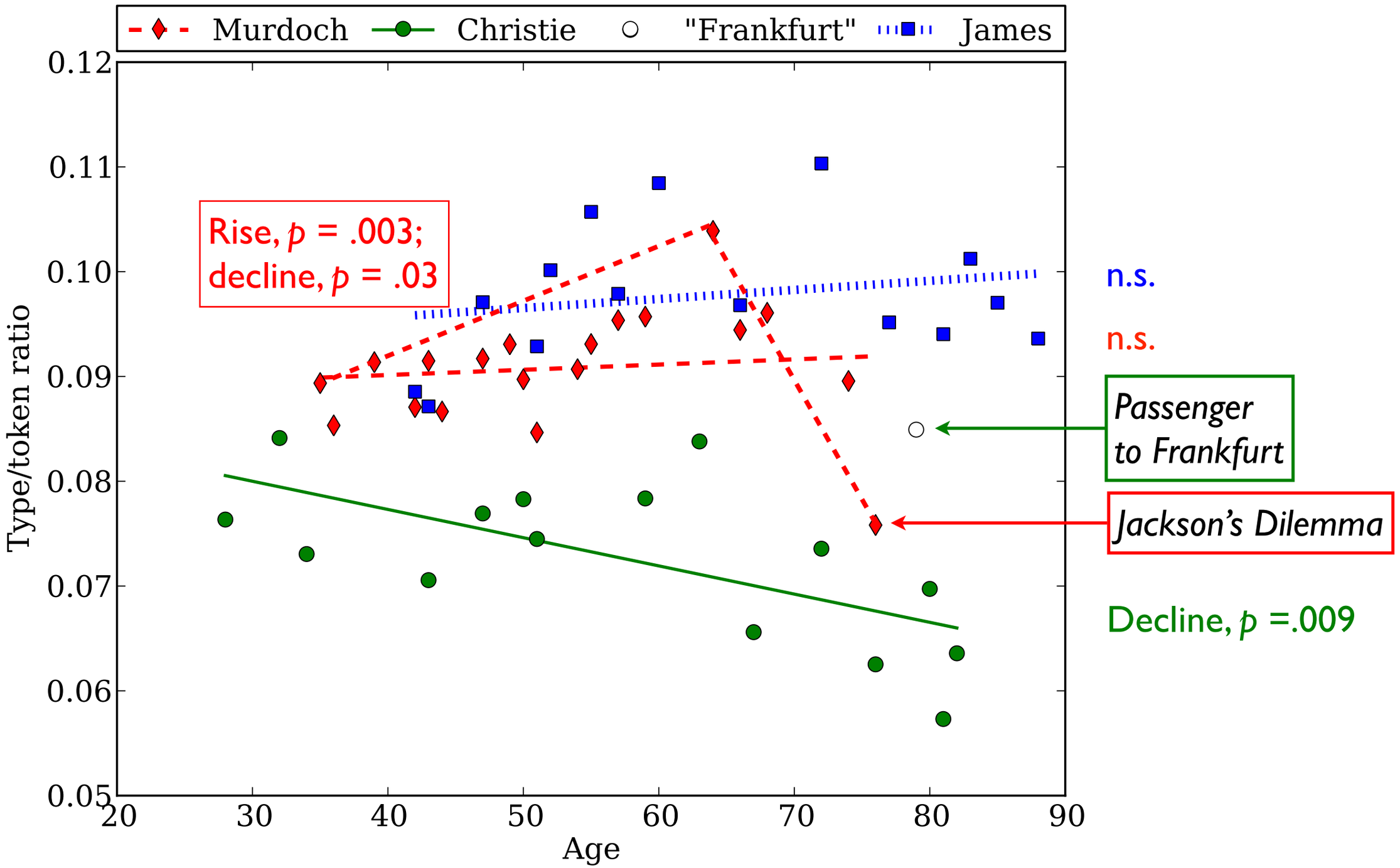
- Find sentence boundaries.
- Generate parse trees (including part-of-speech tagging).
  Charniak (2006) parser.
- No attempt to detect or exclude dialog.
  Very difficult in practice.

5
Measures and Results

Vocabulary size

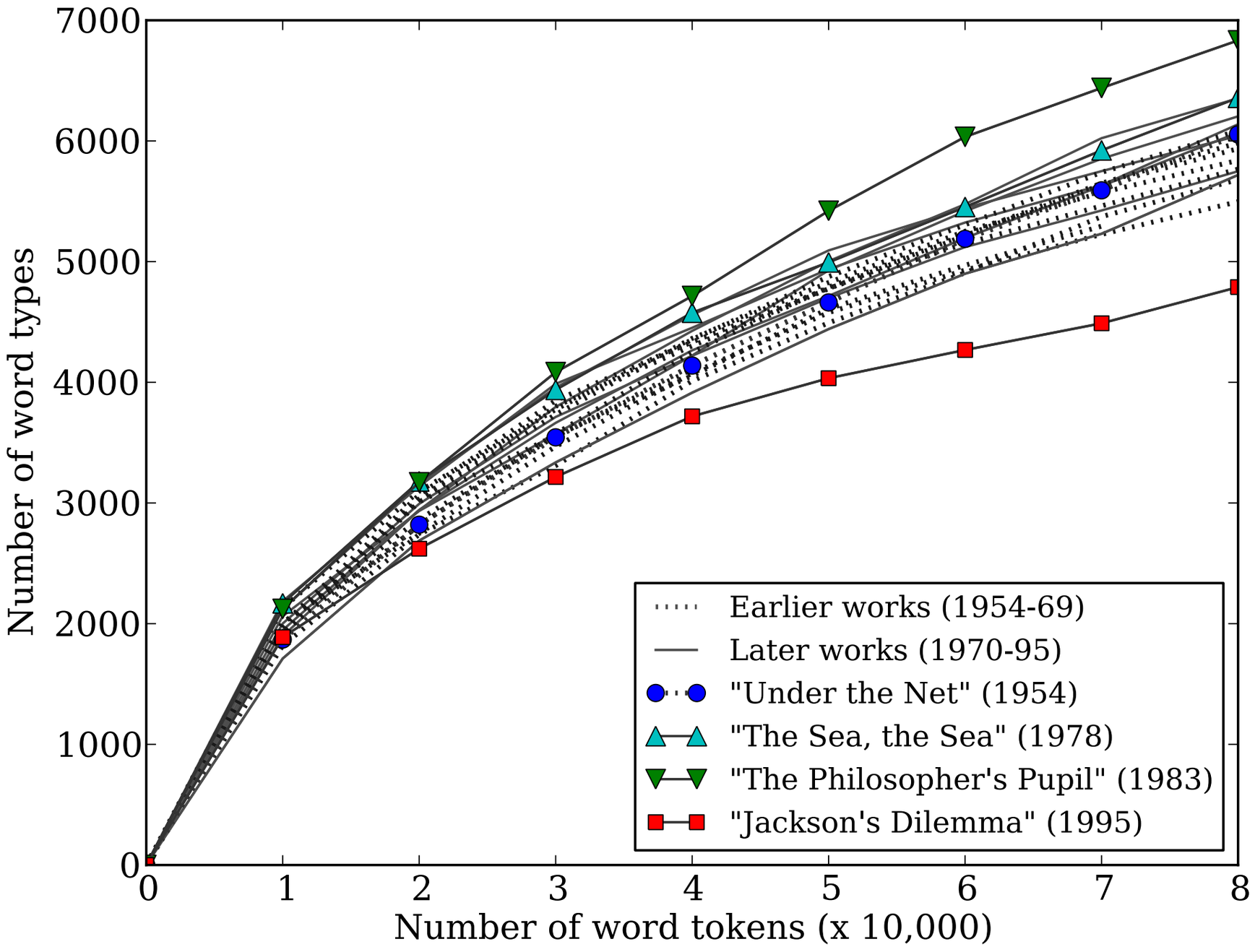
- Type–token ratio
- Word-type introduction rate

Type-token ratio



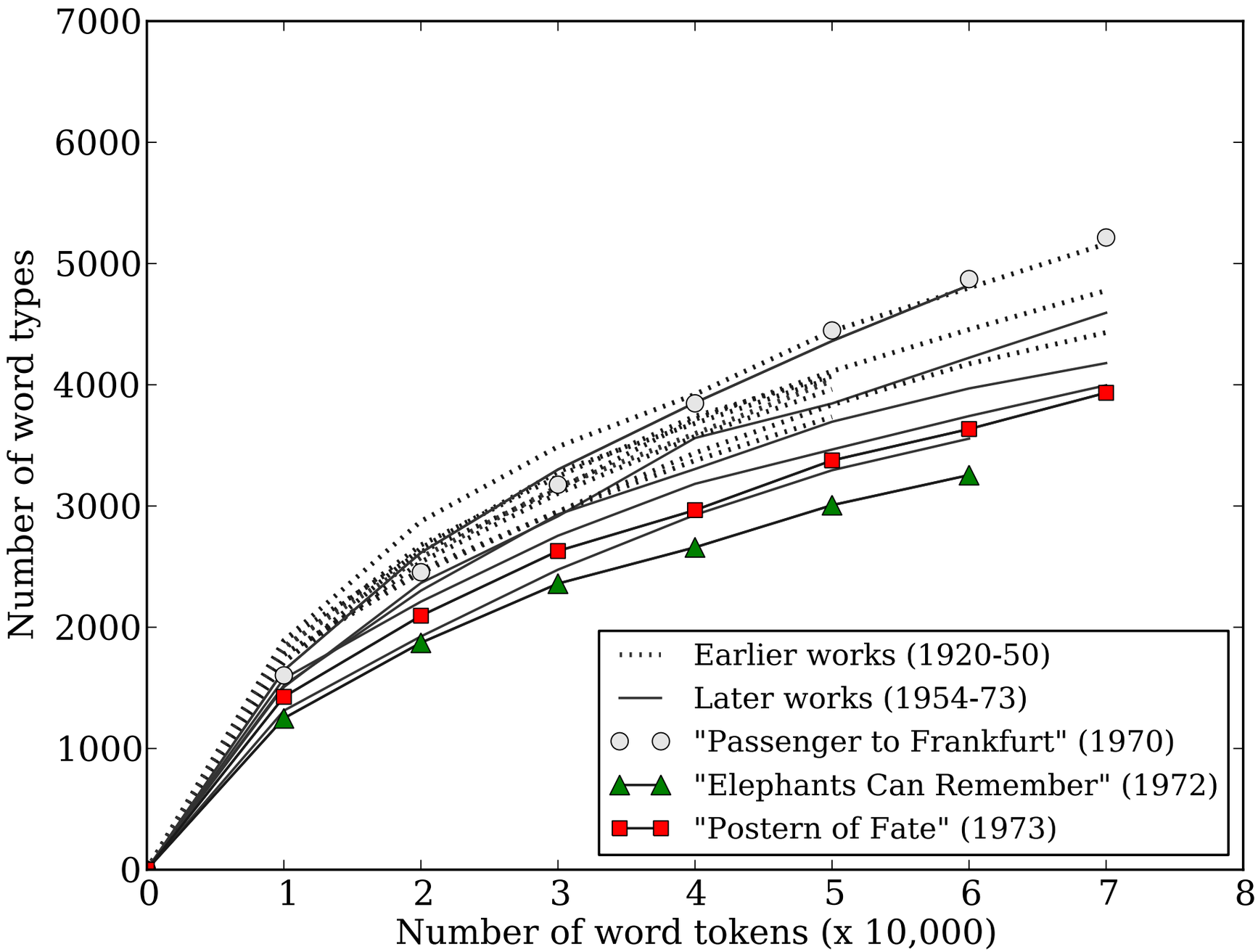
Word-type introduction rate

Iris Murdoch



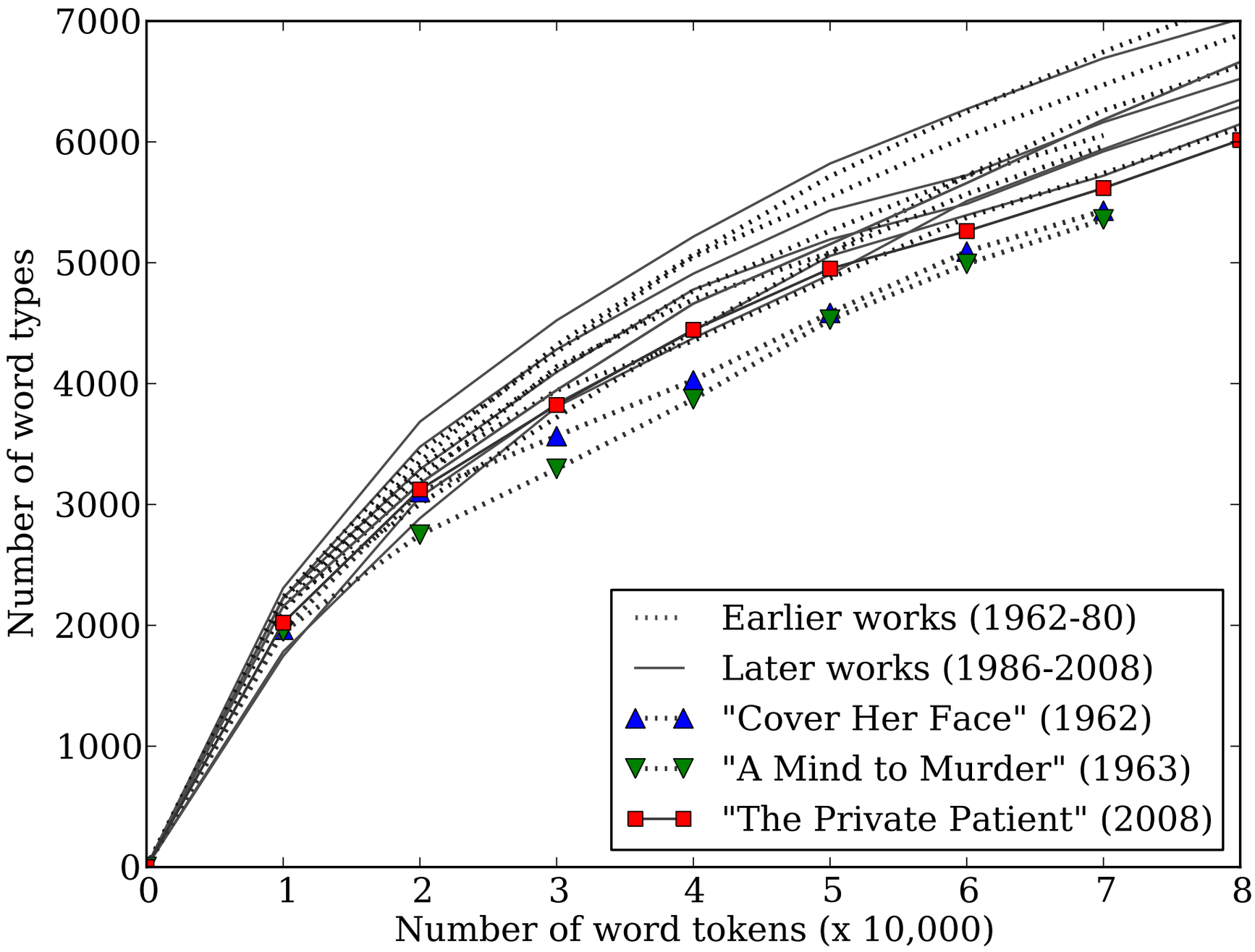
Word-type introduction rate

Agatha Christie



Word-type introduction rate

P.D. James



Repetition

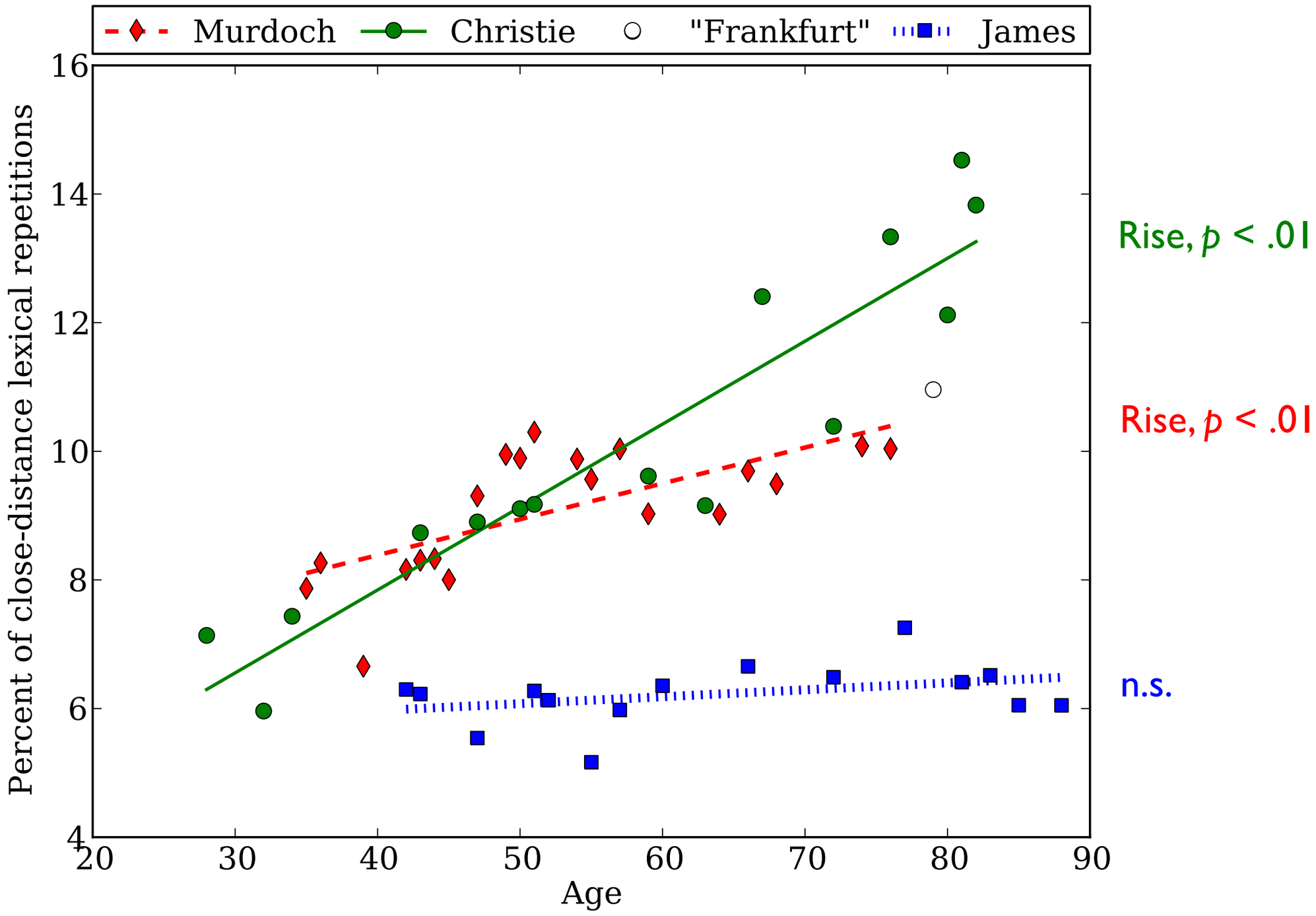
- Repetition of a content word in next 10 content tokens.
- Phrasal (multi-word) repetitions.

Close lexical repetition

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

Agatha Christie, *Postern of Fate* [her final novel]

Lexical repetitions



Rise, $p < .01$

Rise, $p < .01$

n.s.

Phrasal repetition

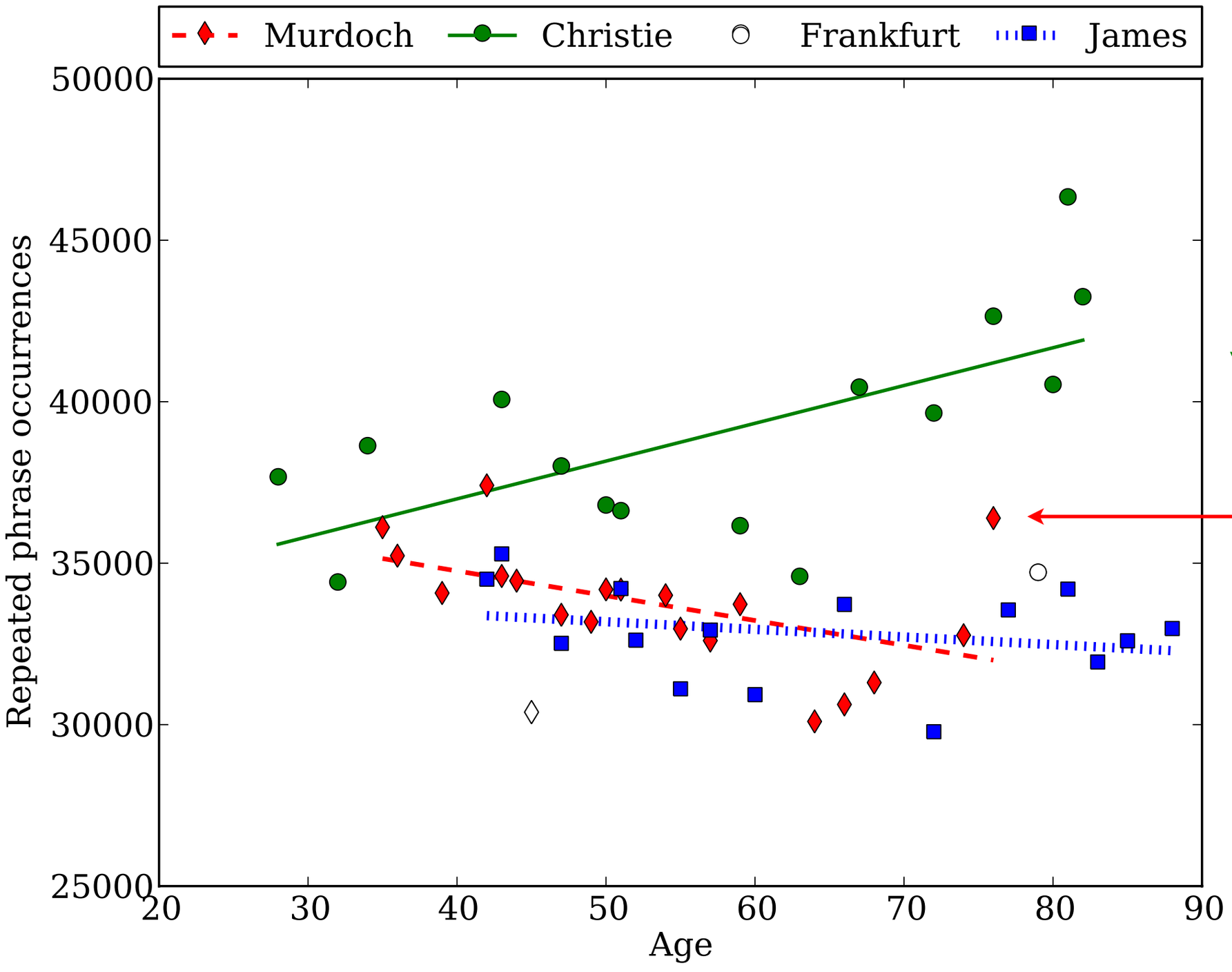
She got near the door. She stopped suddenly, then walked on. It looked as though something like a bundle of clothes was lying near the door. Something they'd pulled out of Mathilde and not thought to look at, Tuppence wondered. She quickened her pace, almost running. When she got near the door, she stopped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

Agatha Christie, *Postern of Fate* [her final novel]

- This passage also contains many repetitions of phrases from elsewhere in the text.

Phrasal repetitions

in first 55K words of each novel



Rise, $p < .01$;
with *Frankfurt*,
rise, $p < .05$

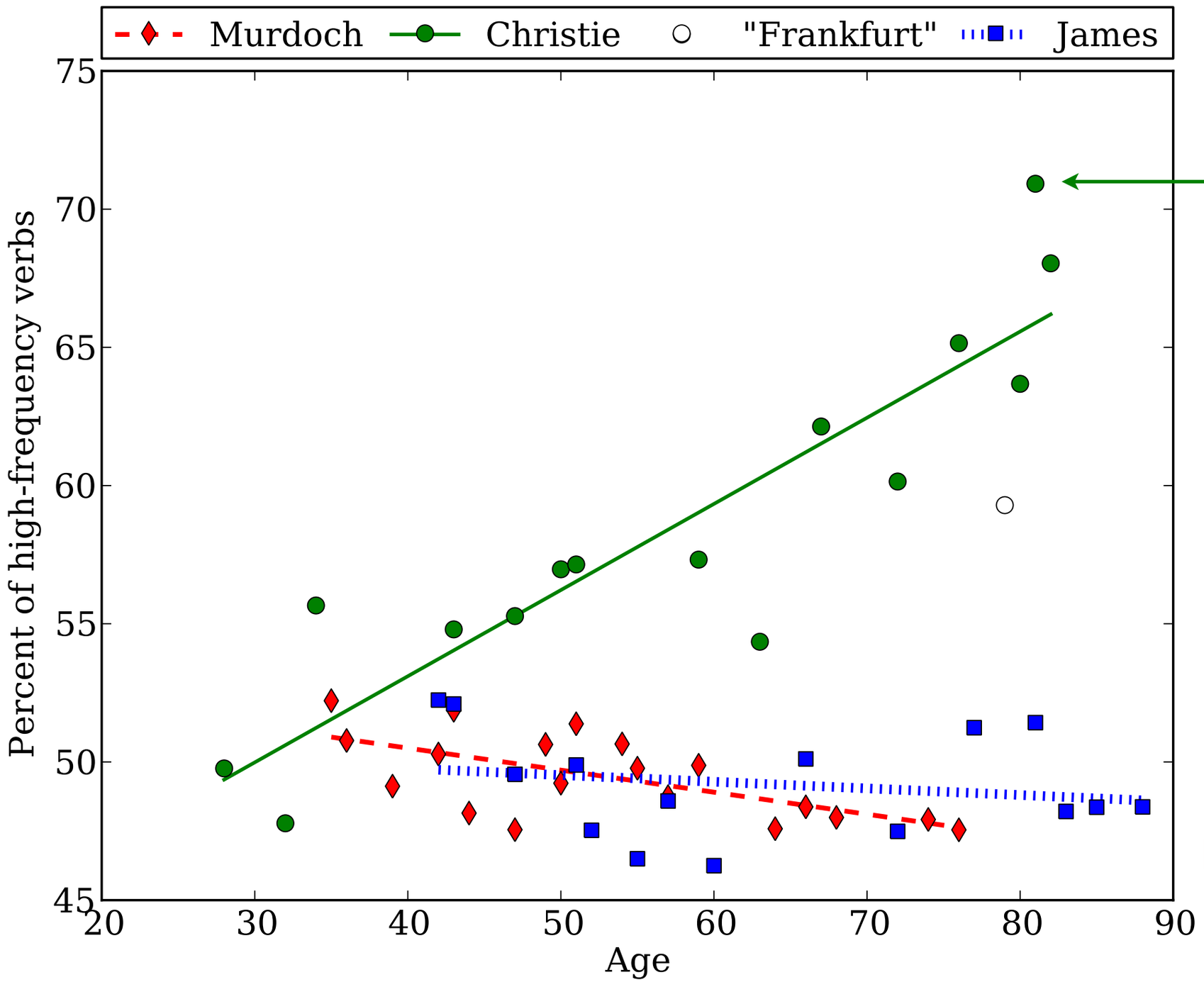
Jackson's Dilemma

n.s.
Decline, $p < .05$

Lexical specificity

- Use of 35 high-frequency unspecific verbs.
  be, come, do, get, give, go, have, know, have, …
- Use of *thing*-words.
  thing, something, anything, nothing

High-frequency verbs



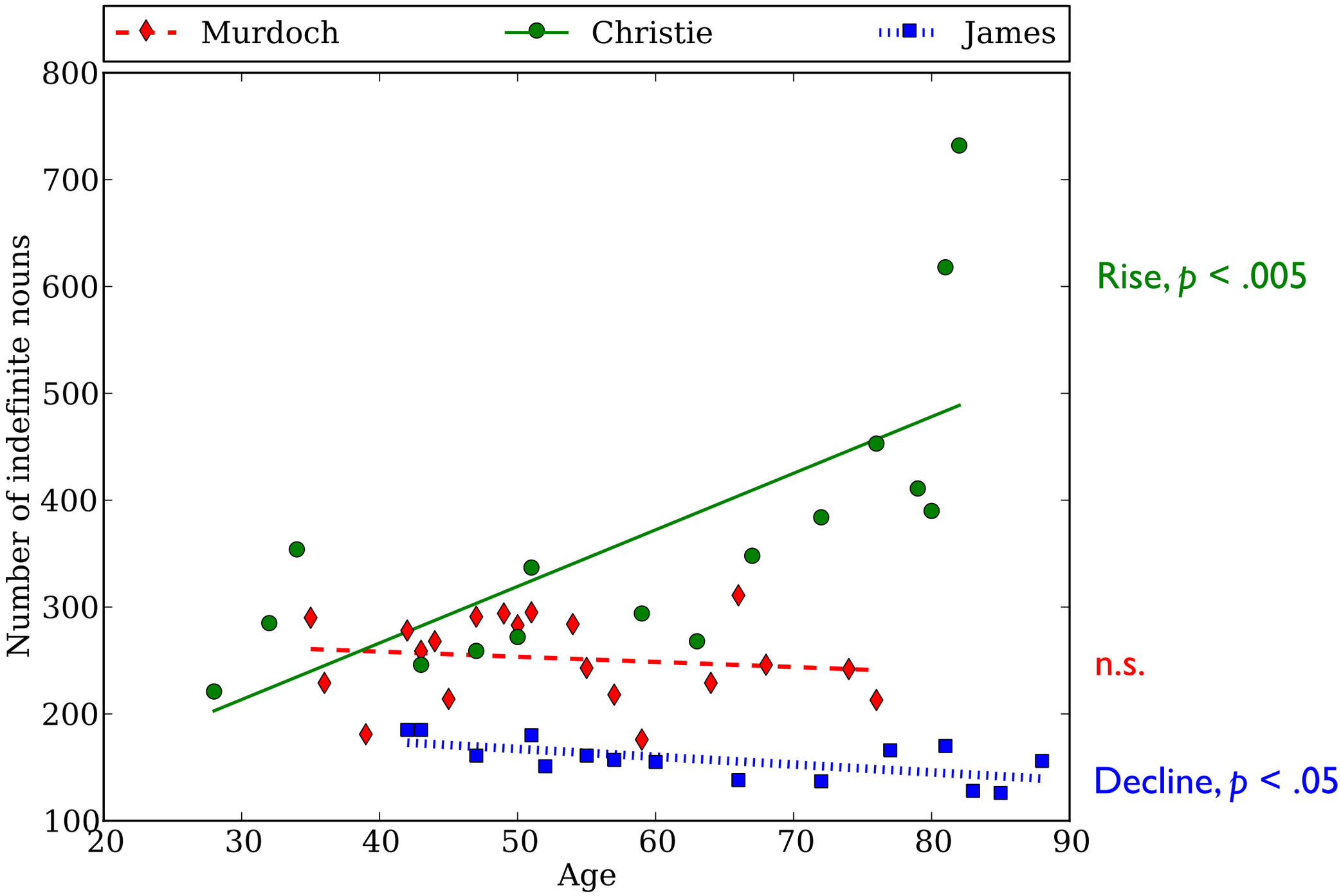
71% of verb tokens accounted for by only 35 verb types.

Rise, $p < .0001$

n.s.

Decline, $p < .003$

Thing-words in first 55K words of each novel



Rise, $p < .005$

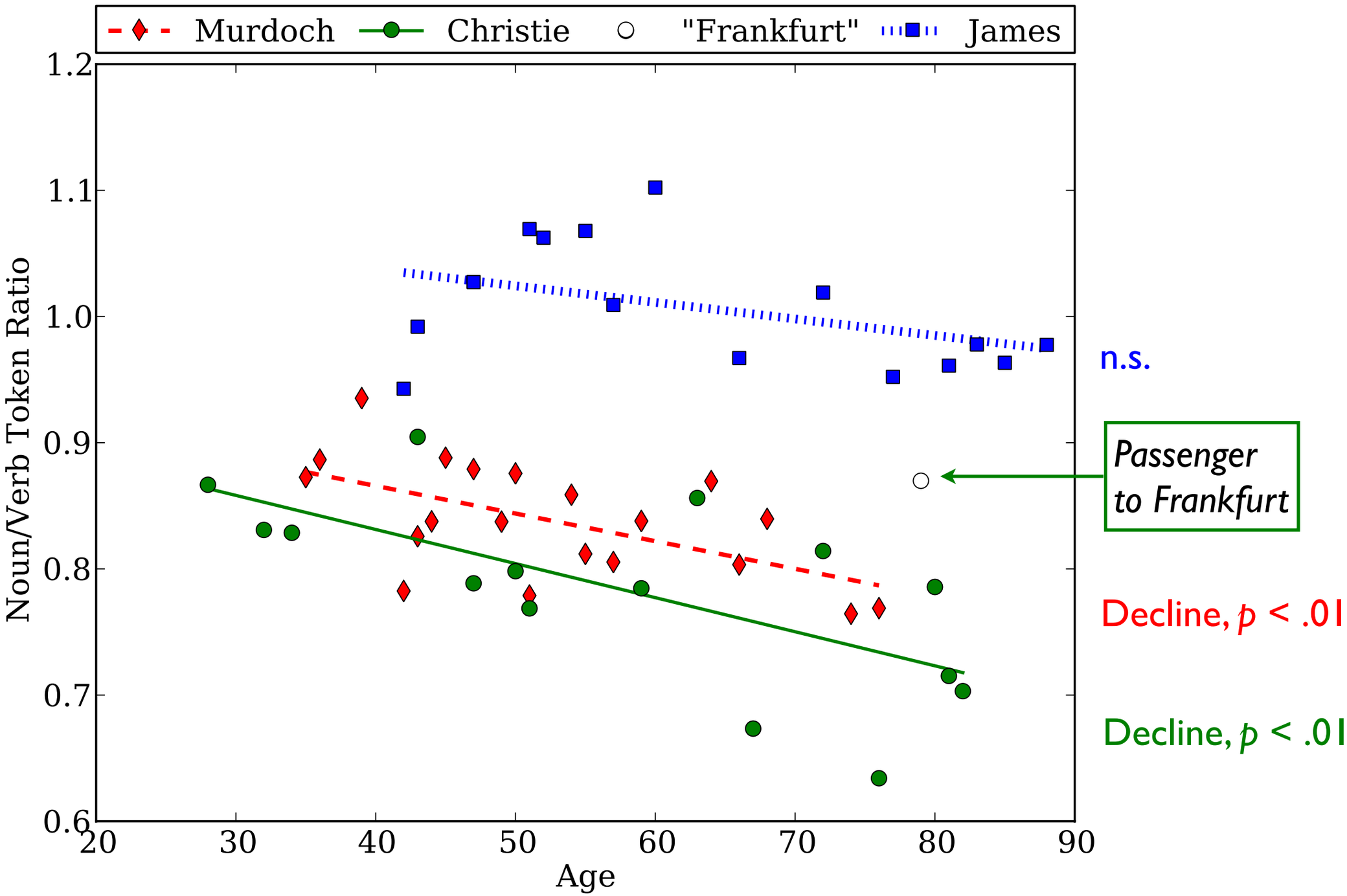
n.s.

Decline, $p < .05$

Word class distribution

- Relative proportions of:
  - (non-proper) nouns
  - content verbs
- … by token count and by types.

Noun-token / verb-token ratio



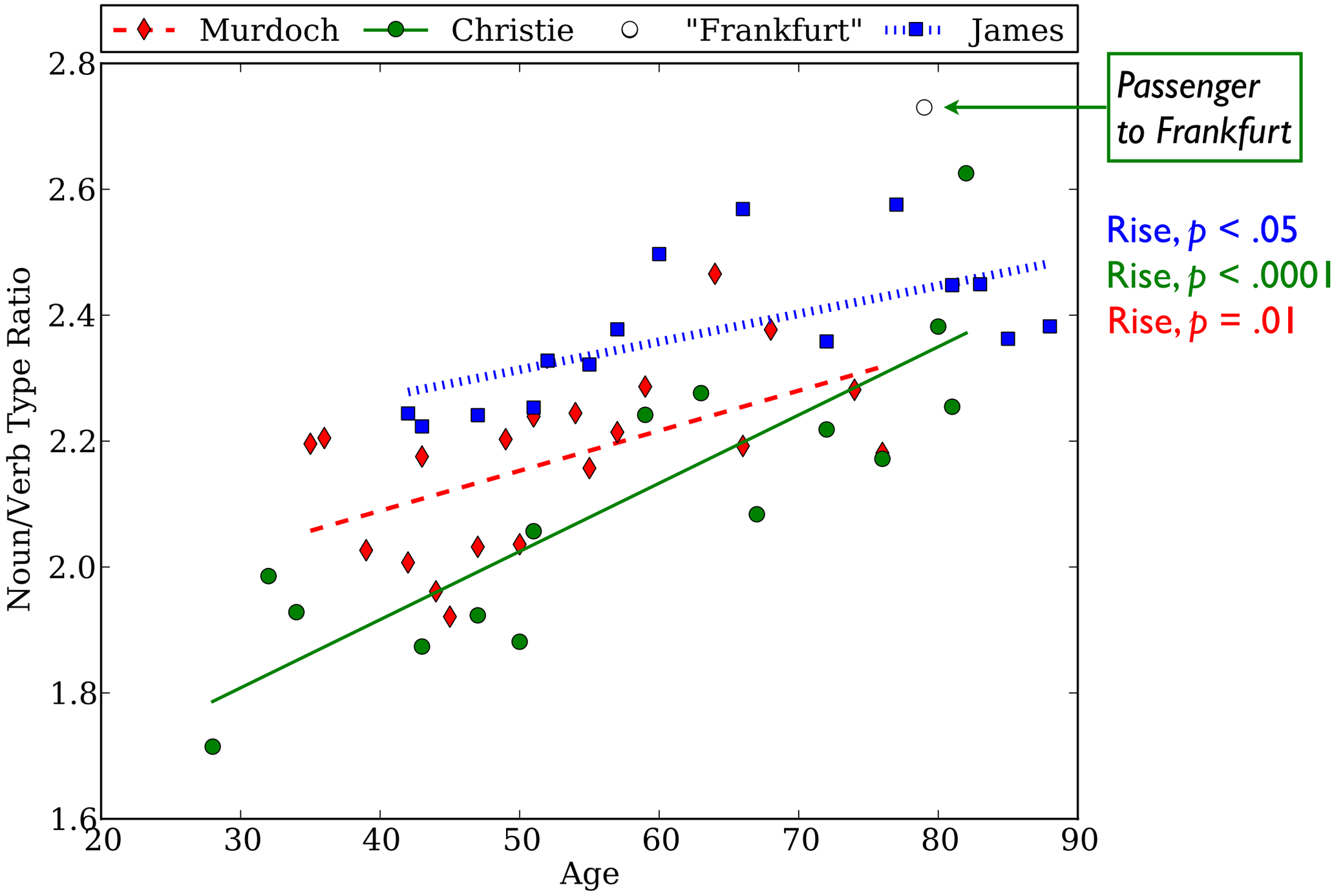
n.s.

Passenger to Frankfurt

Decline, $p < .01$

Decline, $p < .01$

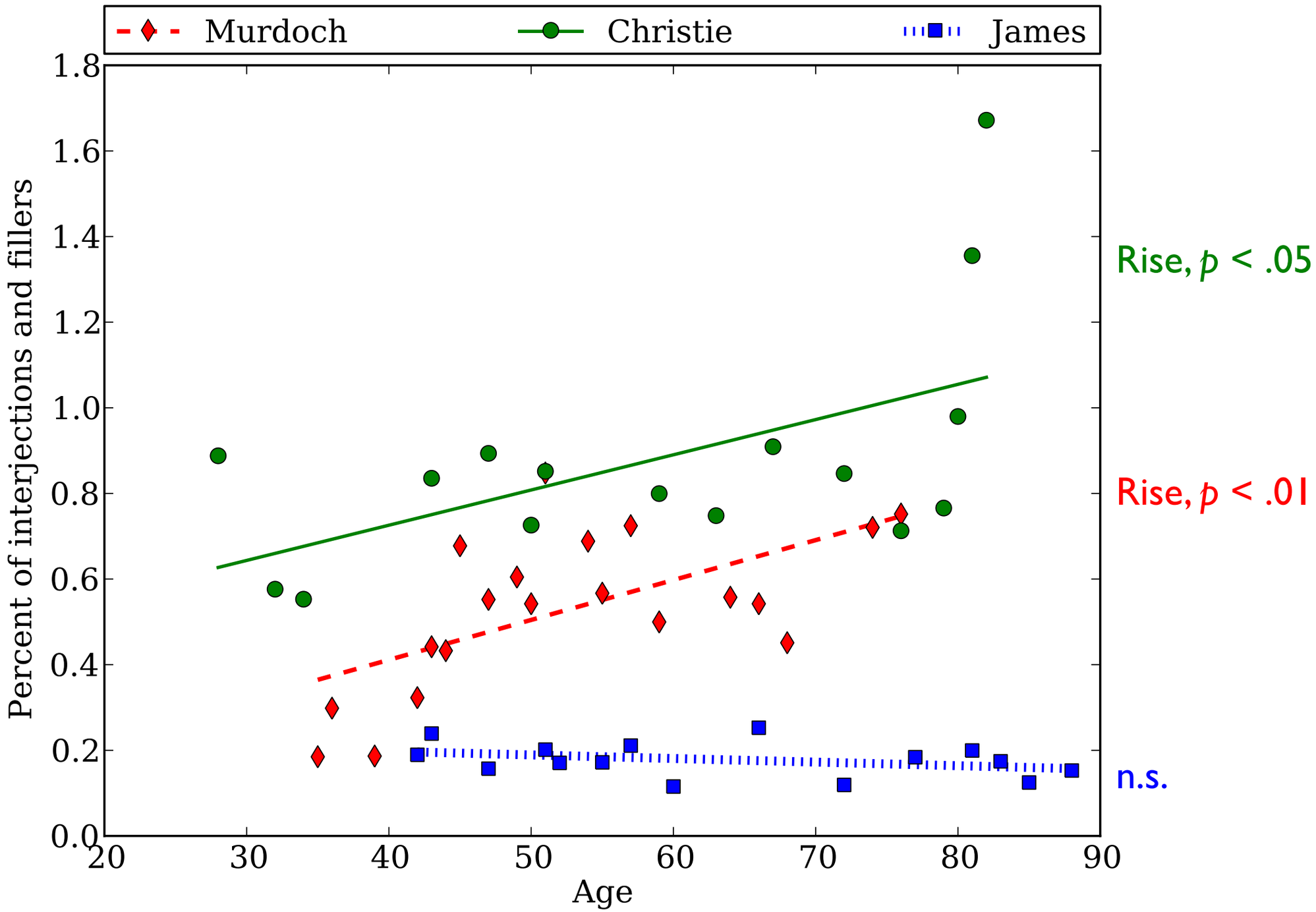
Noun-type / verb-type ratio



Fillers

- Proportion of interjections and fillers (*well, yeah, um, ah*).
- Largely in characters' dialog.
- Assumed to reflect author's idea of natural dialog.

Fillers



Rise, $p < .05$

Rise, $p < .01$

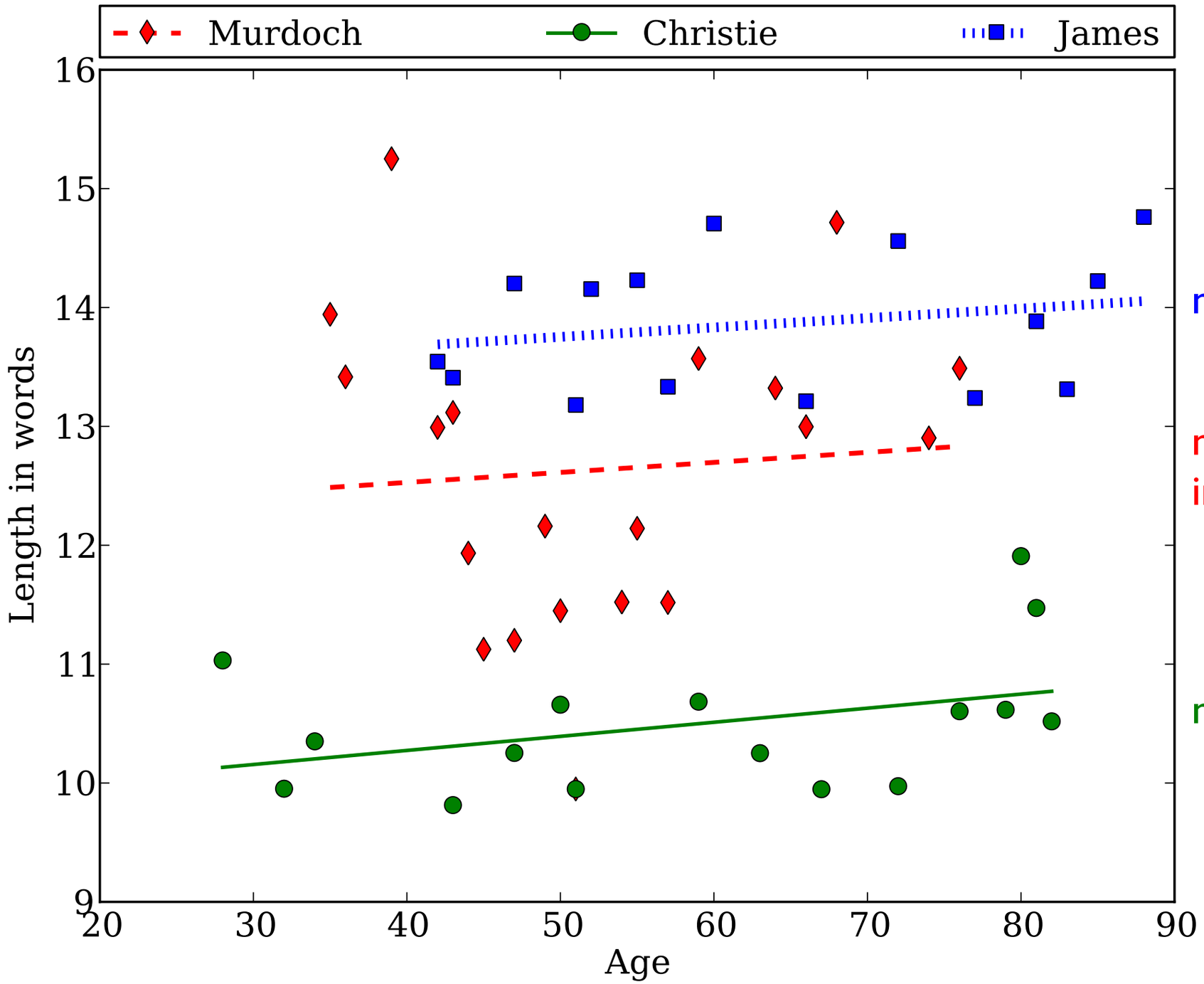
n.s.

Syntactic complexity

- Mean sentence length in words;
  mean number of clauses per sentence.
- Depth of parse tree.
- D-level — qualitative levels of complexity.

- Cheung, H. and Kemper, S. Competing complexity metrics and adults' production of complex sentences. *Applied Psycholinguistics*, 13:53–76, 1992.
- Covington, M.A., *et al*. How complex is that sentence? A proposed revision of the Rosenberg and Abbeduto D-Level Scale. Research Report 2006-01, CASPR, University of Georgia, 2006.
- Rosenberg, S. and Abbeduto, L. Indicators of linguistic competence in the peer group conversational behavior of mildly retarded adults. *Applied Psycholinguistics*, 8:19–32, 1987.

Mean sentence length

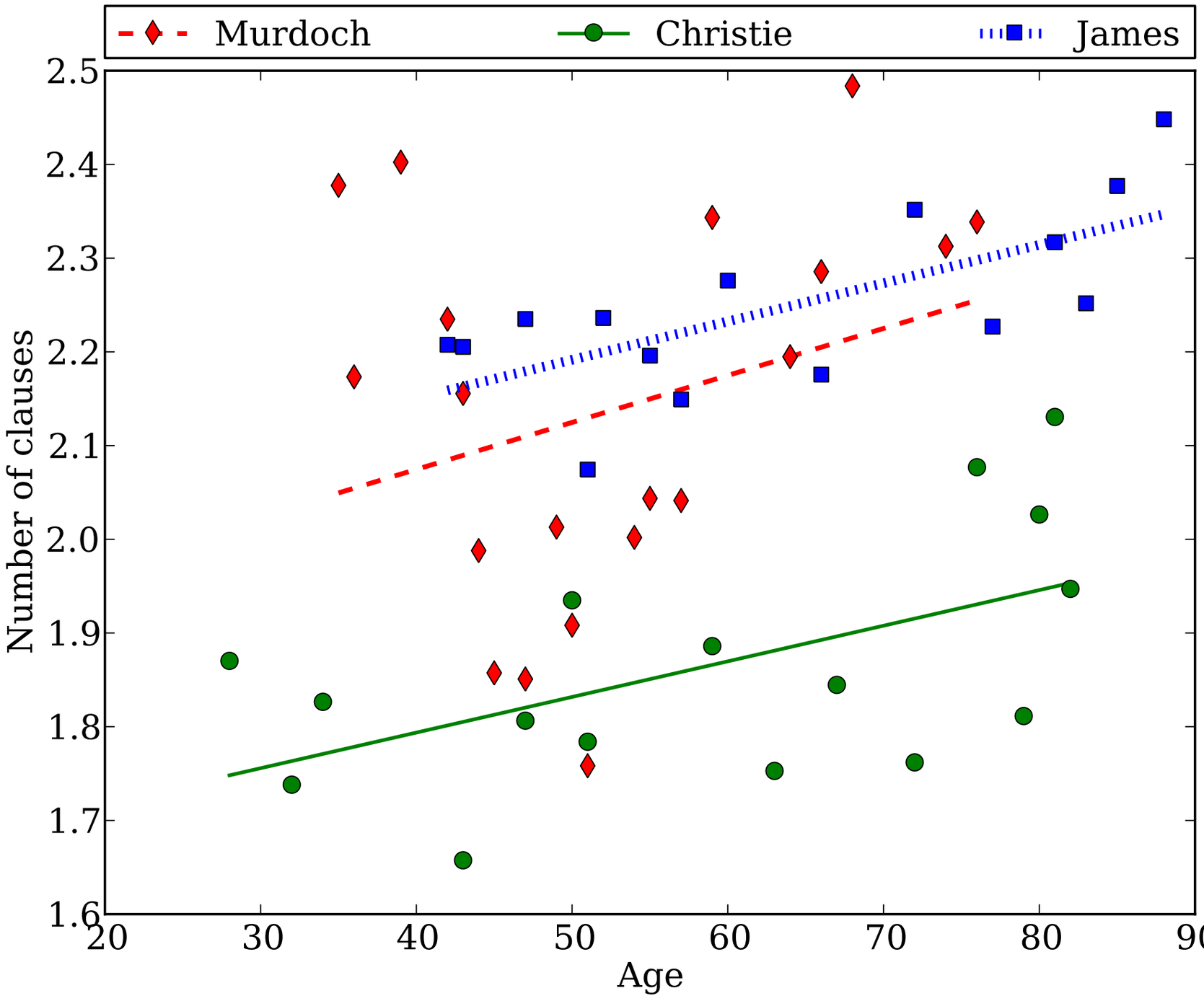


n.s.

n.s.; deep trough in 40s–50s

n.s.

Mean clauses per sentence

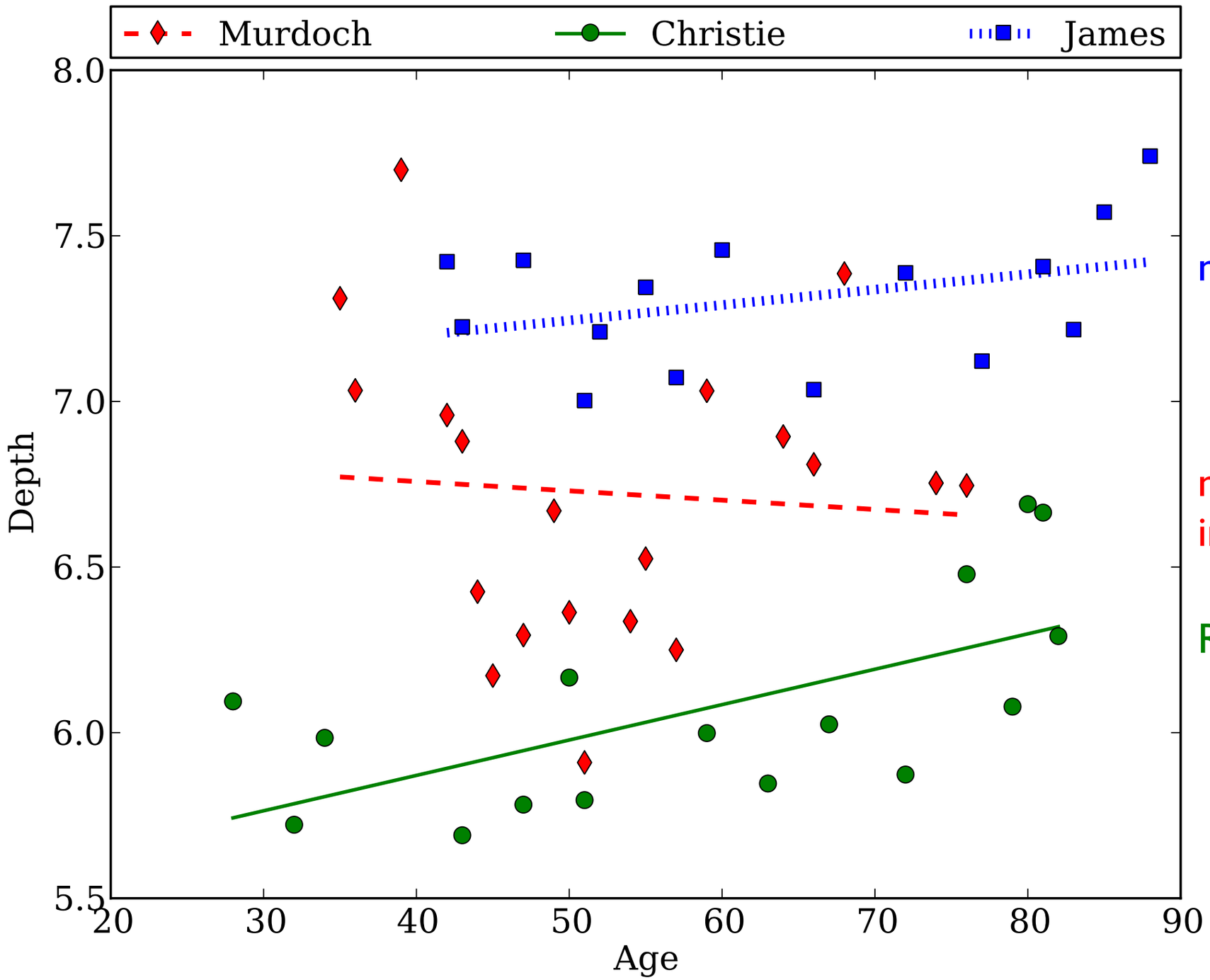


Rise, $p < .01$

Rise, but $p > .05$;
deep trough in 40s–50s

Rise, $p < .05$

Parse tree depth

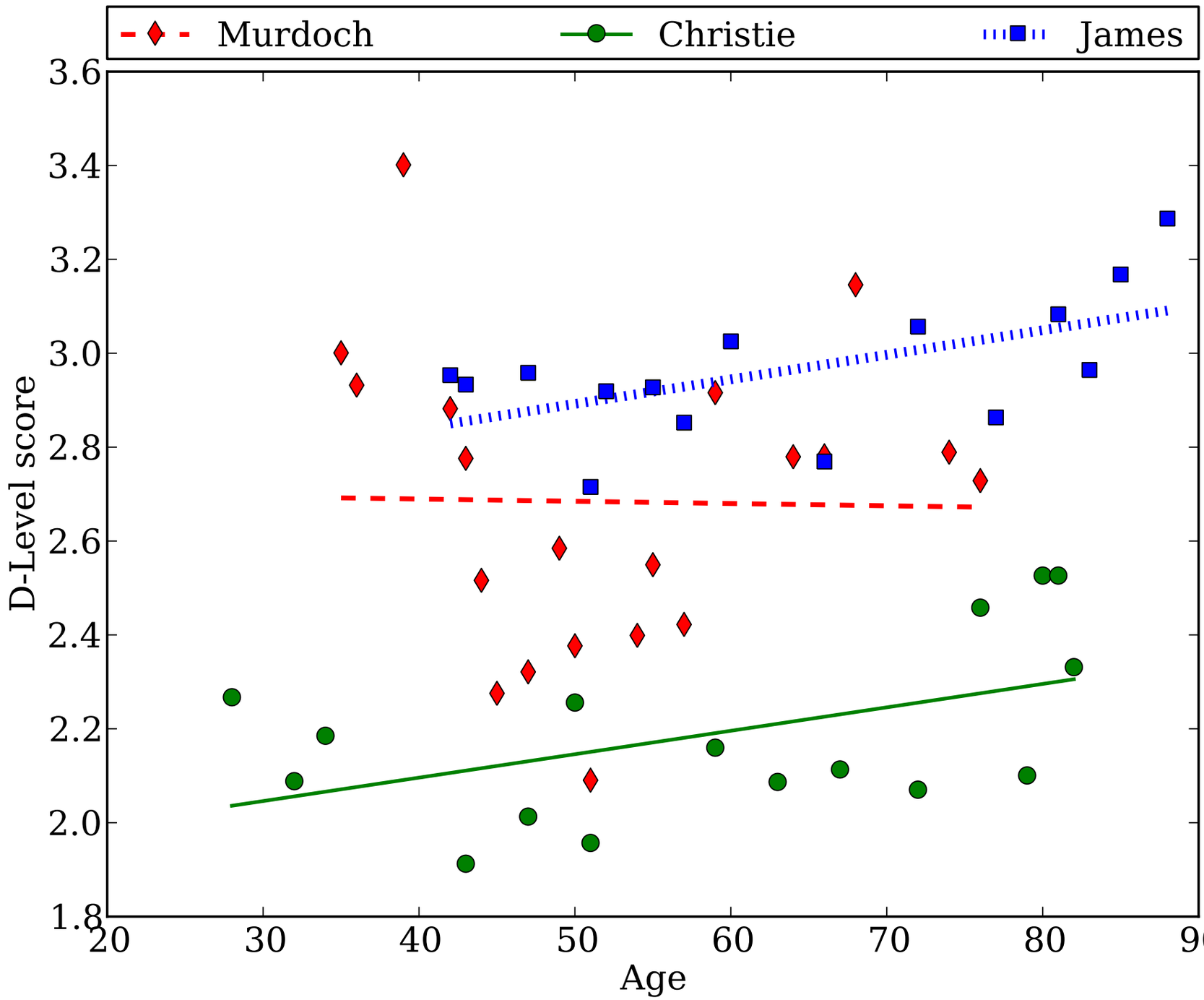


n.s.

n.s.; deep trough in 40s–50s

Rise, $p < .01$

D-level

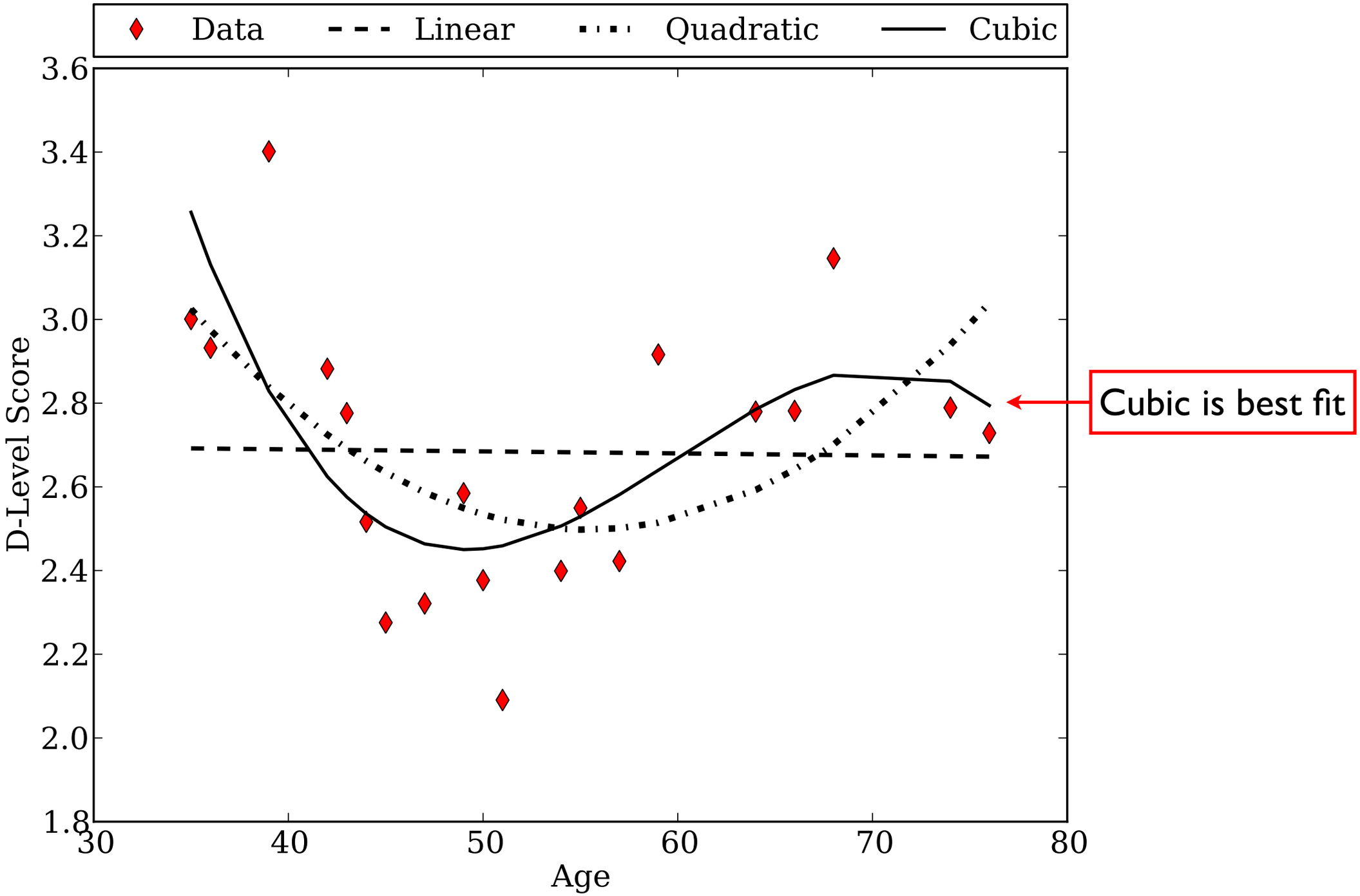


Rise, $p < .03$

n.s.; deep trough in 40s–50s

Rise, $p = .053$

Iris Murdoch's D-level



Passive voice

- Proportion of sentences with passive verb.
- Consider form of auxiliary and presence of agent.

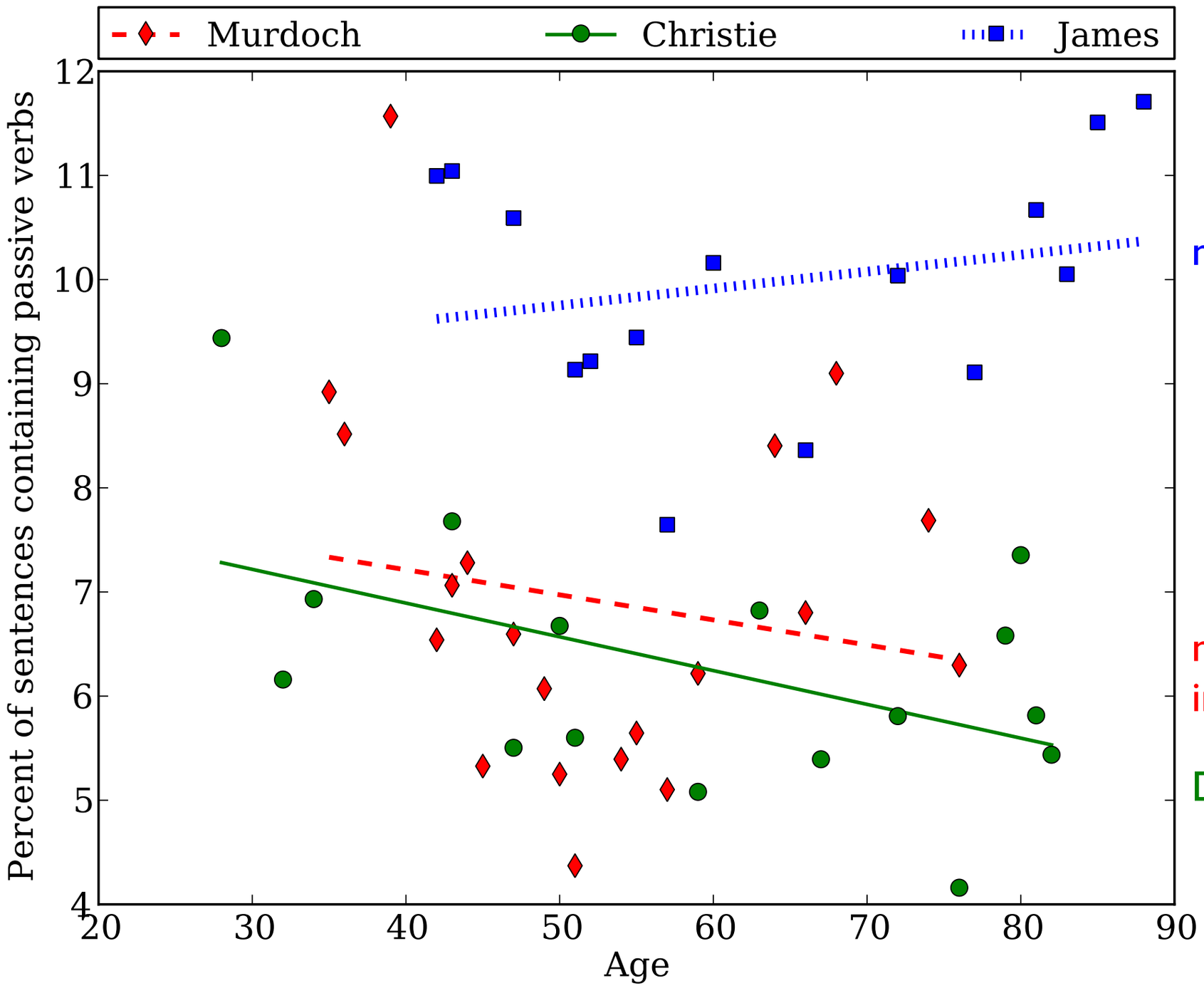
The vase was / got broken in the move. No agent

The vase was / got broken by John. ← Agent

- Ambiguity of verb forms precludes perfect identification of all passives.

Bare passives included only if with agent.

Passive sentences

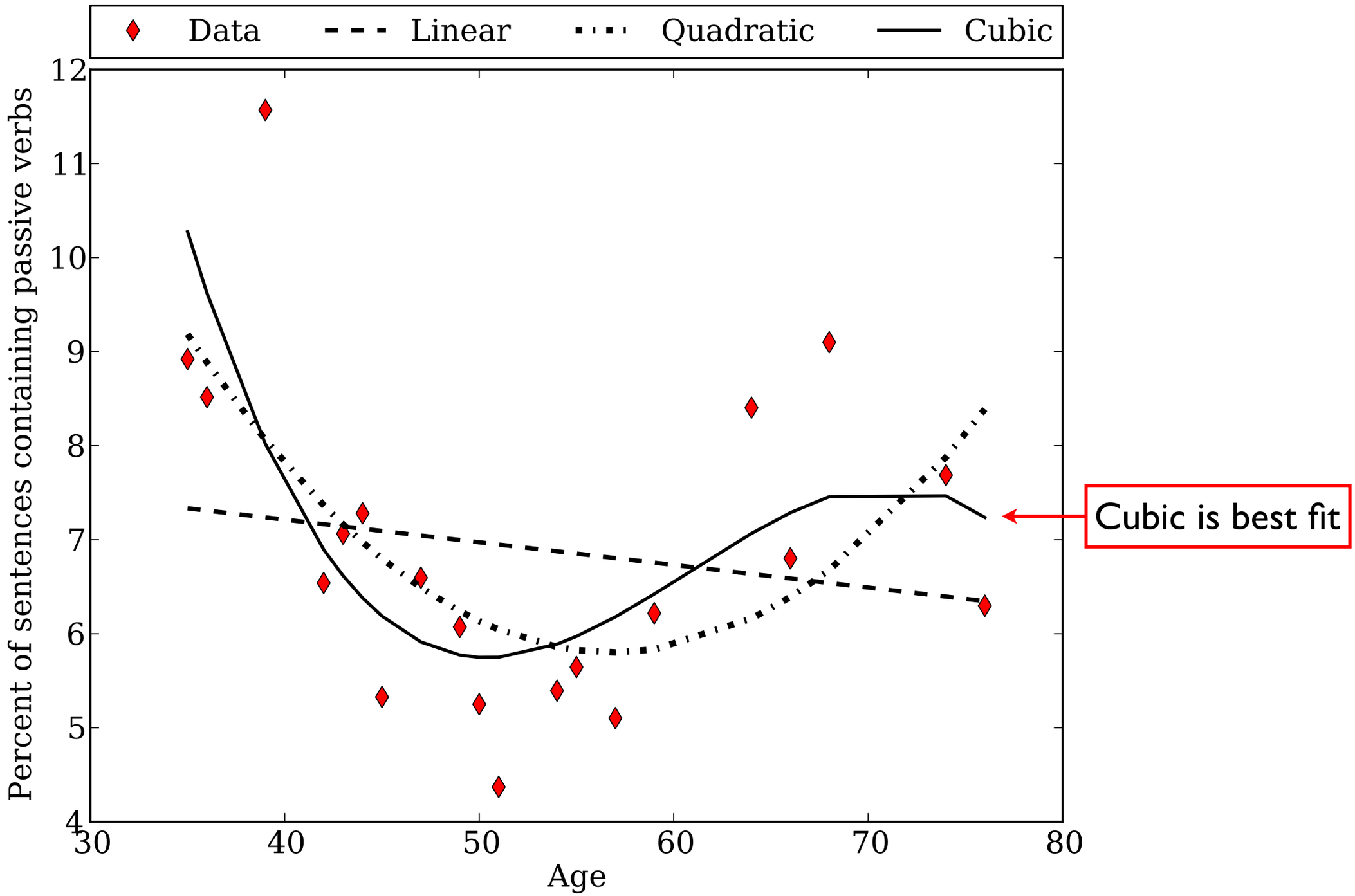


n.s.

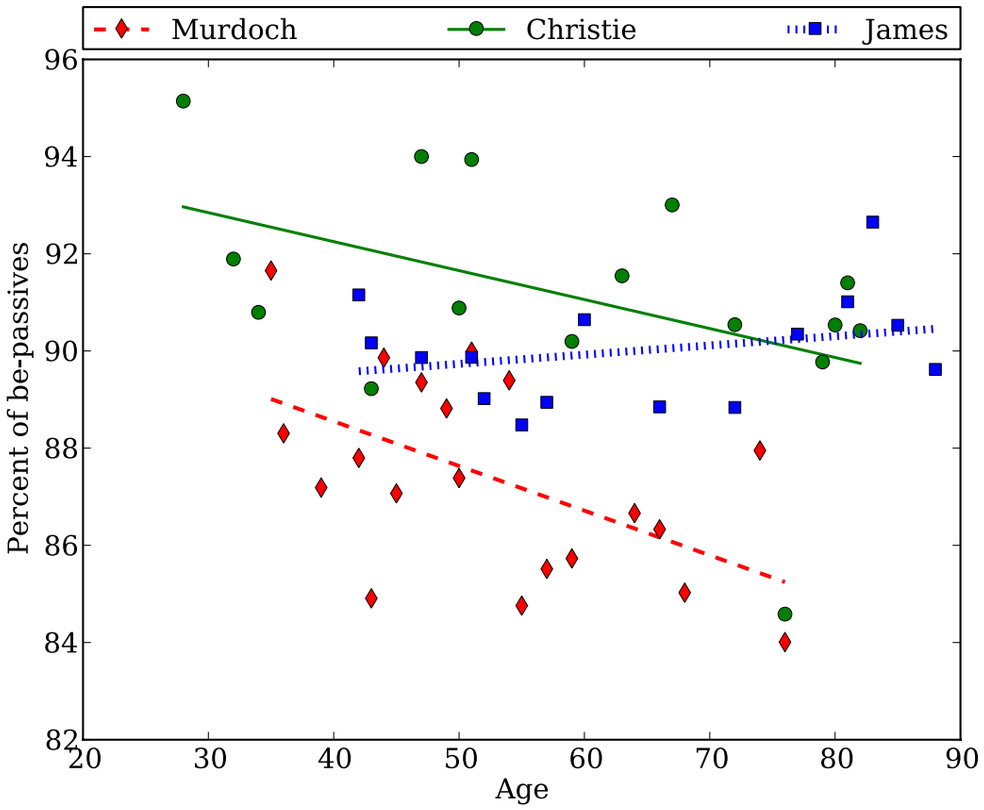
n.s.; deep trough in 40s–50s

Decline, p = .054

Iris Murdoch's passive sentences



Be-passives versus *get*-passives

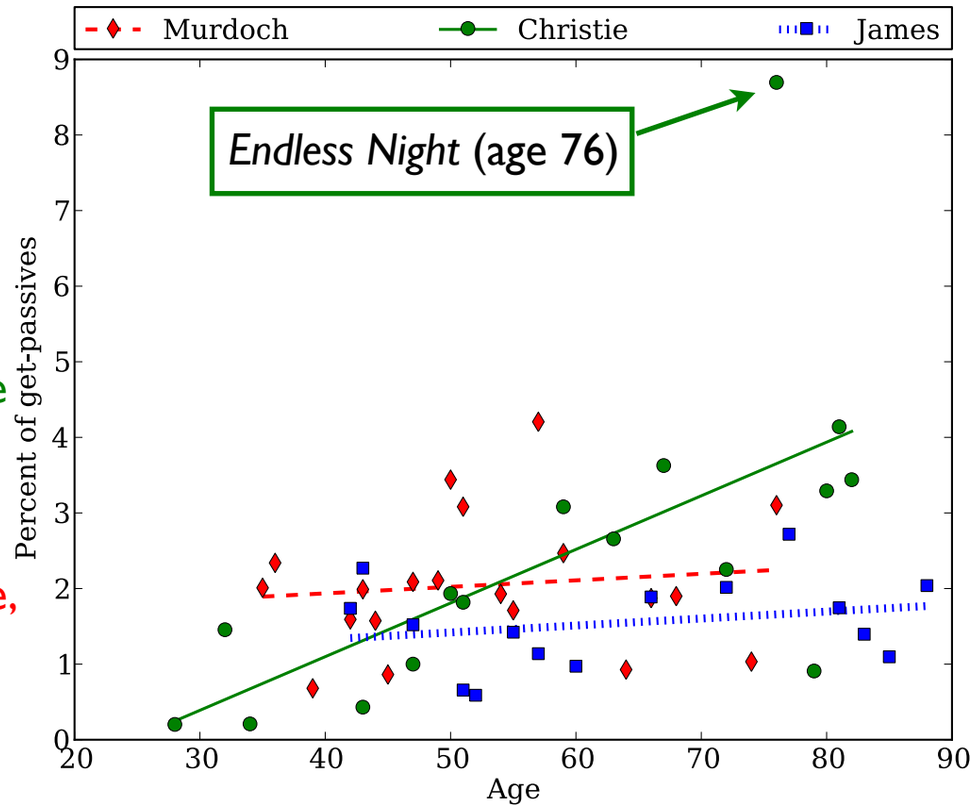


n.s.

Decline

n.s.

Decline,
$p < .05$

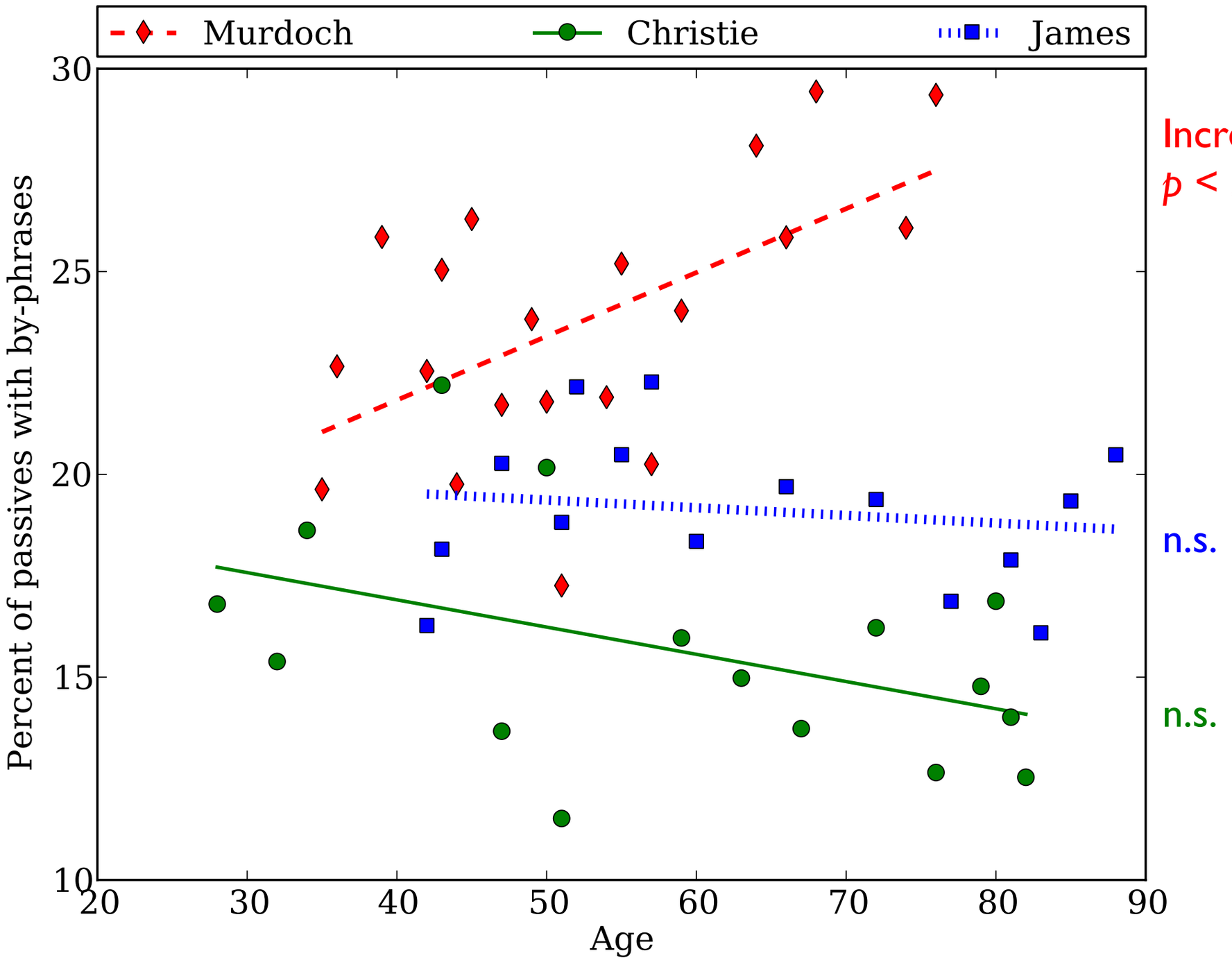


Increase,
$p < .01$

n.s., rise in 50s

n.s.

Passives with *by* agent



Increase, $p < .01$

n.s.

n.s.

6

Summary of results

Hypotheses:

Murdoch and Christie will show changes as in dementia.

James will show little or no change, as in normal aging.

In all cases, we are looking for **relative change** within an individual, not at absolute numbers.

Lexical changes

	Marker	Murdoch	Christie	James
✓	Vocabulary size	Sharp decline in last novel; signs in her 50s	Gradual decline, sharp decline later	No change
✓	Lexical repetition	Increase with sharp rise in 50s	Pronounced increase	No change
	Phrasal repetition	✗ Decline	✓ Pronounced increase	No change
	Word specificity	✗ Decline in high-freq verbs, no change in *thing*-words	✓ Sharp increase in high-freq verbs and *thing*-words	Fewer *thing*-words
✓	Word class distribution	Fewer noun tokens, more verb tokens	Fewer noun tokens, more verb tokens	More verb tokens
✗		Fewer verb types	More noun types, fewer verb types	More noun types, fewer verb types
✓	Fillers	Increase with sharp rise in 50s	Sharp increase	No change

Syntactic changes

	Marker	Murdoch	Christie	James
✗	Syntactic complexity	Irregular; deep decline in 50s	No change	Increase in two measures
	Use of passive voice	?? Sharp drop in 50s	✗ No change	No change
	be-passives	✓ Decrease	✗ No change	No change
	get-passives	?? Rise in 50s	✓ Increase	No change
	Agentless passives	✓ Decrease	✗ No change	No change

Interpretation

- Murdoch shows many signs of AD, but not all that we expected.
  - Not in word specificity or repetition.
  - Unclear in syntactic complexity.
- James shows no signs of AD.
- Christie shows many clear lexical signs of AD, but almost no syntactic signs.

Iris Murdoch's 'trough'

- Drop in syntactic complexity and passive, ~45–60.
- A (more?) troubled period for her.

  "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."
  Iris Murdoch, 26 July 1970 (age 51); quoted by Conradi 2001

- No explanation for recovery around 60.

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

7
Future work

Fill in some gaps

- Look for changes in word-type frequency.
- Look for changes in word specificity.
- Factor phrase length into a repetitiveness index.

Semantics and cohesion

- Null result for changes in propositional density.
  Thanks to Vanessa Feng.
- Look for changes in semantic and in discourse-level cohesion.

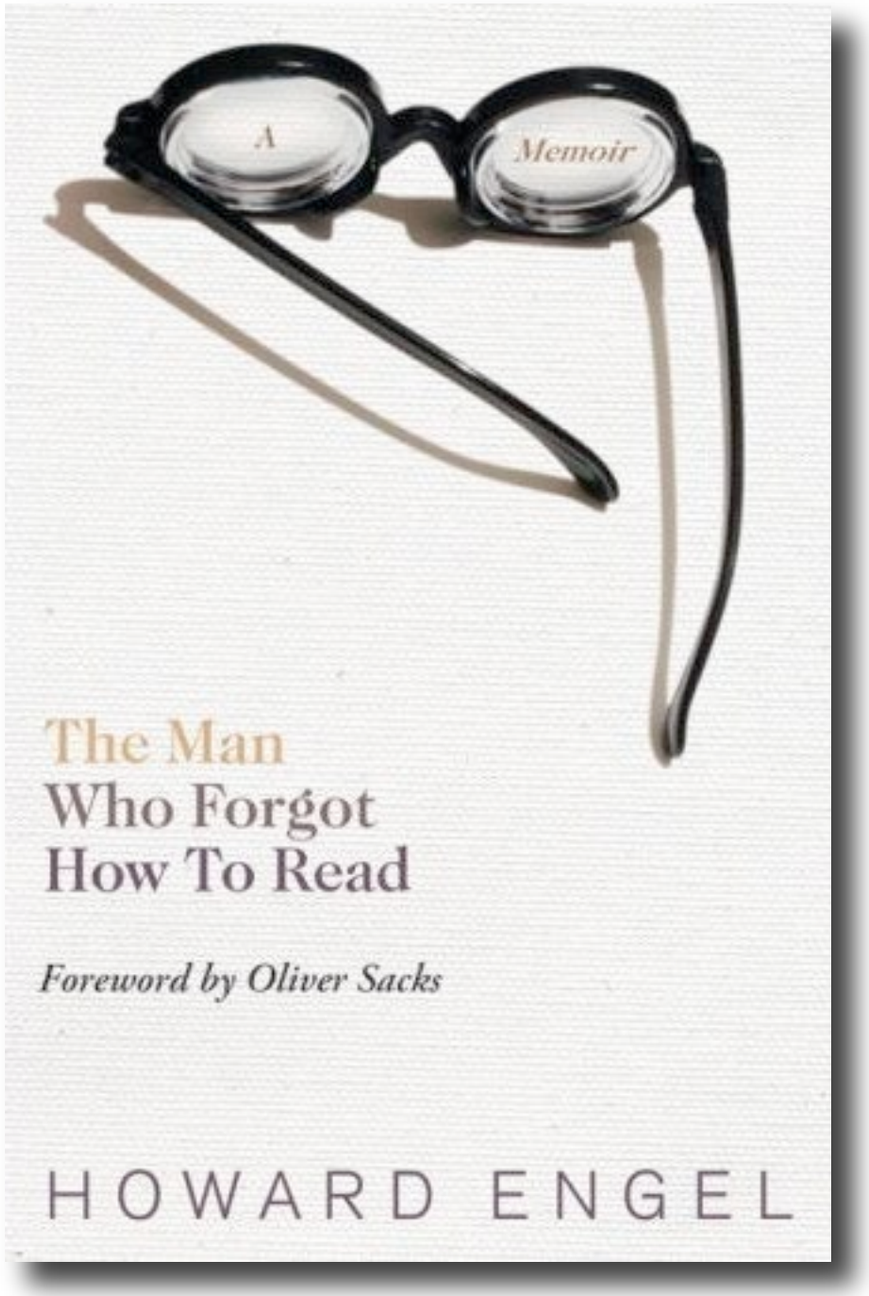
Clinical data

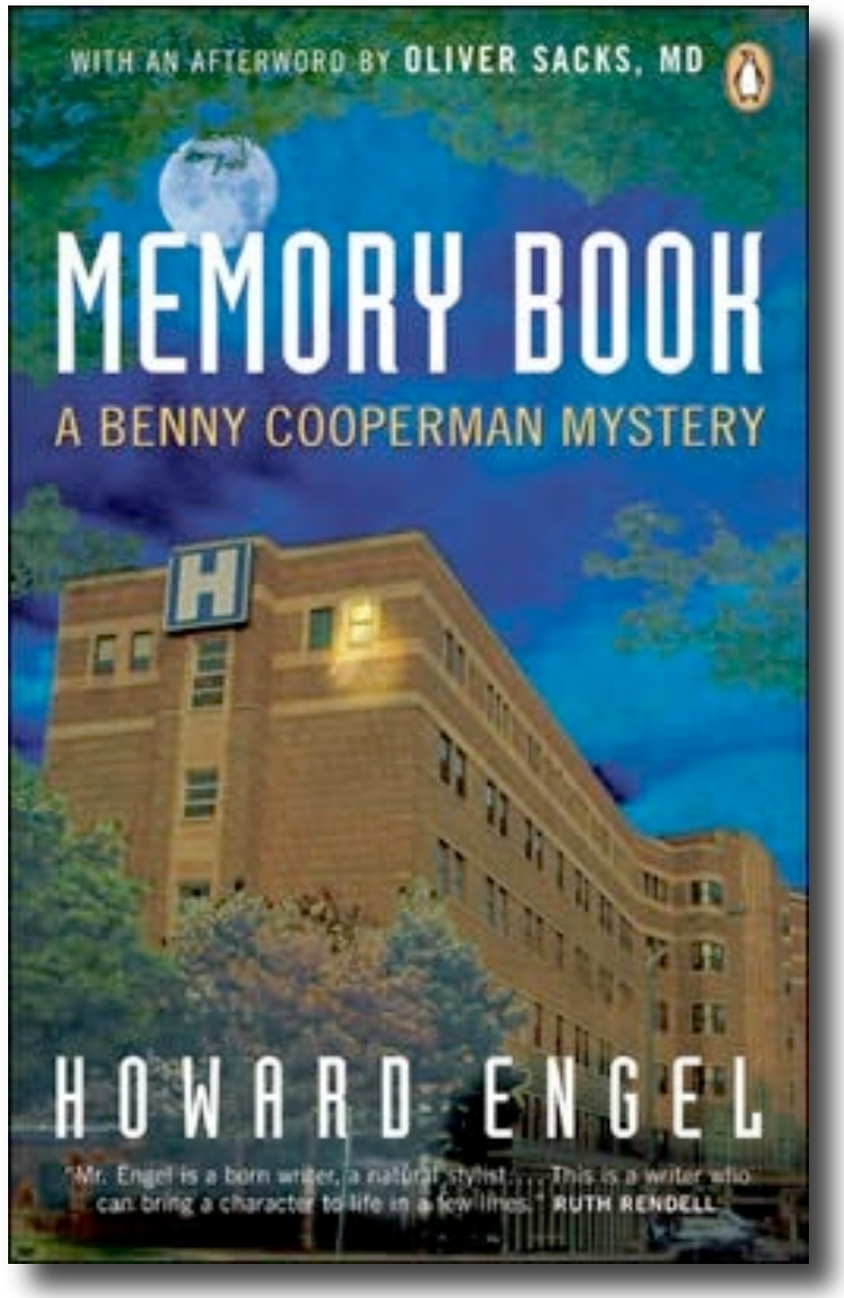
- Transcripts of interviews with Alzheimer's patients.
- Some longitudinal data (from time of diagnosis) available.
- Need to find real patients with long pre-diagnosis text archive (and matched controls).

Alexia sine agraphia

- Loss of ability to read but not to write.
- Howard Engel, Toronto detective novelist.

Photo © Joshua Sherurcij, used by permission


A
Memoir
The Man
Who Forgot
How To Read
Foreword by Oliver Sacks
HOWARD ENGEL


WITH AN AFTERWORD BY OLIVER SACKS, MD
MEMORY BOOK
A BENNY COOPERMAN MYSTERY
HOWARD ENGEL
"Mr. Engel is a born writer, a natural stylist. . . . This is a writer who can bring a character to life in a few lines." RUTH RENDELL

7
Conclusion

Conclusion

- Evidence that warning signs of Alzheimer's dementia can be detected in writing.
- Should your word processor be looking out for you?
  - Difficult issues in health communication and ethics.

Thank you

Removing dialog

- Ideally, treat dialog and narrative separately.
- Naive in-out algorithm vulnerable to error.
- OCR problems despite error checking.
- Interleaved speech and narrative.
- Other uses of quotation marks.