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
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

- No changes in discourse competence. Better in synonyms test.
- But overall, differences are small.

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.

2

Language in Alzheimer’s disease
Alzheimer’s dementia

• Most common form of dementia.
• Caused by Alzheimer’s disease.
  – Cortical degeneration.

Alzheimer cells

Amyloid plaques

Neurofibrillary tangles

Healthy cells

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, …

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.

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?!

# Lexical changes

<table>
<thead>
<tr>
<th>Marker</th>
<th>Dementia</th>
<th>Healthy aging</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vocabulary size</td>
<td>Sharp decrease</td>
<td>Gradual increase, then possible slight decrease</td>
</tr>
<tr>
<td>Lexical repetition</td>
<td>Pronounced increase</td>
<td>Possible small change</td>
</tr>
<tr>
<td>Word specificity</td>
<td>Pronounced decrease</td>
<td>Possible small change</td>
</tr>
<tr>
<td>Word class distribution</td>
<td>Fewer nouns, compensation in verbs</td>
<td>No change</td>
</tr>
<tr>
<td>Fillers</td>
<td>Pronounced increase</td>
<td>Possible slight increase</td>
</tr>
</tbody>
</table>
## Syntactic changes

<table>
<thead>
<tr>
<th>Marker</th>
<th>Dementia</th>
<th>Healthy aging</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syntactic complexity</td>
<td>Sharp decline</td>
<td>Little or no change, then possible rapid decline in mid-70s</td>
</tr>
<tr>
<td>Use of passive voice</td>
<td>Pronounced decrease</td>
<td>Possible small decrease</td>
</tr>
<tr>
<td>Auxiliary verb in passive voice</td>
<td>Get dominates <em>be</em></td>
<td><em>Be</em> dominates <em>get</em></td>
</tr>
<tr>
<td>Passives without agent</td>
<td>Greater decrease</td>
<td>Moderate decrease</td>
</tr>
</tbody>
</table>
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
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.

• 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.

P.D. James, 1920–
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

<table>
<thead>
<tr>
<th>Author</th>
<th>No of novels</th>
<th>Age at composition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Christie</td>
<td>16</td>
<td>28 to 82</td>
</tr>
<tr>
<td>Murdoch</td>
<td>20</td>
<td>35 to 76</td>
</tr>
<tr>
<td>James</td>
<td>15</td>
<td>42 to 88</td>
</tr>
</tbody>
</table>
• 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).
- 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

Rise, $p = .003$; decline, $p = .03$

Decline, $p = .009$
Word-type introduction rate
Iris Murdoch

![Graph showing the word-type introduction rate for Iris Murdoch's works. The graph plots the number of word types against the number of word tokens (x 10,000). The data is broken down into earlier works (1954-69), later works (1970-95), and specific novels such as "Under the Net" (1954), "The Sea, the Sea" (1978), "The Philosopher's Pupil" (1983), and "Jackson's Dilemma" (1995).]
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.
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

Rise, $p < .0001$

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

Decline, $p < .003$

n.s.
**Thing-words** in first 55K words of each novel

![Graph showing the number of indefinite nouns against age for Murdoch, Christie, and James.](image)

- **Murdoch**:
  - Rise, $p < .005$

- **Christie**:
  - n.s.

- **James**:
  - 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

Passenger to Frankfurt

Decline, \( p < .01 \)

n.s.
Noun-type / verb-type ratio

Rise, $p < .0001$
Rise, $p < .05$
Rise, $p = .01$

Passenger to Frankfurt
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, $p < .01$

Rise, $p < .05$

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.


Mean sentence length

![Graph showing mean sentence length by age for Murdoch, Christie, and James.]

- 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

Rise, $p < .01$

n.s.; deep trough in 40s–50s

n.s.
D-level

Rise, \( p < .03 \)

n.s.; deep trough in 40s–50s

Rise, \( p = .053 \)
Iris Murdoch’s D-level Cubic is best fit
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.; deep trough in 40s–50s

Decline, \( p = .054 \)
Iris Murdoch’s passive sentences

Cubic is best fit
Be-passives versus get-passives

- Murdoch
- Christie
- James

Endless Night (age 76)

n.s., rise in 50s

Increase, $p < .01$

Decline, $p < .05$

n.s.
Passives with *by* agent

Increase, \( p < .01 \)
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

<table>
<thead>
<tr>
<th>Marker</th>
<th>Murdoch</th>
<th>Christie</th>
<th>James</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓ Vocabulary size</td>
<td>Sharp decline in last novel; signs in her 50s</td>
<td>Gradual decline, sharp decline later</td>
<td>No change</td>
</tr>
<tr>
<td>✓ Lexical repetition</td>
<td>Increase with sharp rise in 50s</td>
<td>Pronounced increase</td>
<td>No change</td>
</tr>
<tr>
<td>✓ Phrasal repetition</td>
<td>✗ Decline</td>
<td>✓ Pronounced increase</td>
<td>No change</td>
</tr>
<tr>
<td>✓ Word specificity</td>
<td>✗ Decline in high-freq verbs, no change in <em>thing</em>-words</td>
<td>✓ Sharp increase in high-freq verbs and <em>thing</em>-words</td>
<td>Fewer <em>thing</em>-words</td>
</tr>
<tr>
<td>✓ Word class distribution</td>
<td>Fewer noun tokens, more verb tokens</td>
<td>Fewer noun tokens, more verb tokens</td>
<td>More verb tokens</td>
</tr>
<tr>
<td>✗ Fewer verb types</td>
<td></td>
<td>More noun types, fewer verb types</td>
<td>More noun types, fewer verb types</td>
</tr>
<tr>
<td>✓ Fillers</td>
<td>Increase with sharp rise in 50s</td>
<td>Sharp increase</td>
<td>No change</td>
</tr>
</tbody>
</table>
## Syntactic changes

<table>
<thead>
<tr>
<th>Marker</th>
<th>Murdoch</th>
<th>Christie</th>
<th>James</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>✗</strong> Syntactic complexity</td>
<td>Irregular; deep decline in 50s</td>
<td>No change</td>
<td>Increase in two measures</td>
</tr>
<tr>
<td>Use of passive voice</td>
<td>?? Sharp drop in 50s</td>
<td>✗ No change</td>
<td>No change</td>
</tr>
<tr>
<td><strong>✓</strong> be-passives</td>
<td>✗ Decrease</td>
<td>✗ No change</td>
<td>No change</td>
</tr>
<tr>
<td>get-passives</td>
<td>?? Rise in 50s</td>
<td>✓ Increase</td>
<td>No change</td>
</tr>
<tr>
<td><strong>✓</strong> Agentless passives</td>
<td>✗ Decrease</td>
<td>✗ No change</td>
<td>No change</td>
</tr>
</tbody>
</table>
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.

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
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.