Reliable measures of syntactic and lexical complexity: The case of Iris Murdoch

Stefan Evert\textsuperscript{1}, Sebastian Wankerl\textsuperscript{1}, Elmar Nöth\textsuperscript{2}

\textsuperscript{1}Corpus Linguistics Group, \textsuperscript{2}Pattern Recognition Lab
University of Erlangen-Nuremberg
25 July 2017
Introduction (1)

Different types of quantitative measures of complexity like

- vocabulary richness (e.g. type-token ratio (TTR))
- readability (e.g. age-of-acquisition)
- syntactic complexity (e.g. Yngve (1960) depth, sentence length)

play an important role in a wide range of applications, e.g.

- investigating stylometric differences and authorship (Stamatatos 2009)
- studying diachronic changes in grammar (Bentz et al. 2014)
- assessing readability and difficulty level of a text (Graesser et al. 2004)
- exploring the characteristics of translated texts (Volansky, Ordan, and Wintner 2015)
- identifying determinants of style in scientific writing (Bergsma, Post, and Yarowsky 2012)
- multivariate analysis of linguistic variation (Diwersy, Evert, and Neumann 2014)
Introduction (2)

Some researchers have used complexity measures to detect early symptoms of Alzheimer’s disease (AD).

However, as in most other research on complexity measures

• sampling variation is not taken into account
• significance tests are rarely applied
• or inappropriate tests are used
Our Case Study

Previous work on detecting Alzheimer’s disease

- based on the works of British novelist Iris Murdoch (diagnosed with AD)
- using different measures of readability, lexical and syntactic complexity
Our Case Study

Previous work on detecting Alzheimer’s disease

- based on the works of British novelist Iris Murdoch (diagnosed with AD)
- using different measures of readability, lexical and syntactic complexity
- obtained conflicting results
  - decline of complexity in Murdoch’s last novel, for various measures (Garrard et al. 2005; Pakhomov et al. 2011)
  - no clear effects, in particular for syntactic complexity (Le et al. 2011)
Our Case Study

Previous work on detecting Alzheimer’s disease

- based on the works of British novelist Iris Murdoch (diagnosed with AD)
- using different measures of readability, lexical and syntactic complexity
- obtained conflicting results
  - decline of complexity in Murdoch’s last novel, for various measures (Garrard et al. 2005; Pakhomov et al. 2011)
  - no clear effects, in particular for syntactic complexity (Le et al. 2011)
- which can be explained by lack of significance testing
  - no confidence intervals for complexity scores of each novel
  - only Le et al. (2011) applied significance tests, but for linear trends over time (not reasonable considering the progress of the disease)
Research Questions

1. How can appropriate significance tests be implemented?
2. Does AD manifest in a significant decline of complexity?
Research Questions

1. How can appropriate significance tests be implemented?
2. Does AD manifest in a significant decline of complexity?

To achieve this goal, we develop new methodology for computation of

- reliable confidence intervals
- significance tests

by combining ideas from

- bootstrapping (Efron 1979)
- cross-validation
The writings of Iris Murdoch

- renowned British author of the post-war era
- published a total of 26 novels
- mostly well received by literary critics
- last novel received “without enthusiasm” (Garrard et al. 2005), Murdoch experienced unexpected difficulties while composing it
- diagnosis of AD shortly after publication

Hence, early symptoms of AD should be found in this last novel.
The writings of Iris Murdoch

- renowned British author of the post-war era
- published a total of 26 novels
- mostly well received by literary critics
- last novel received “without enthusiasm” (Garrard et al. 2005), Murdoch experienced unexpected difficulties while composing it
- diagnosis of AD shortly after publication

Hence

early symptoms of AD should be found in this last novel

Murdoch novel reveals Alzheimer’s

The last novel by the author Iris Murdoch reveals the first signs of Alzheimer’s disease, experts say.

A team from University College London say their examination of works from throughout Dame Iris’s career could be used to help diagnose others.

They found the structure and grammar of her novels was relatively unchanged, but her language was noticeably simpler in her last novel, ‘Jackson’s Dilemma’.

The study is published online by the journal Brain.
Our Corpus

- 19 of Murdoch’s 26 novels
- including the nine last novels, spanning a period of almost 20 years
- acquired as e-books (no errors due to OCR)

Further preprocessing

- Stanford CoreNLP (Manning et al. 2014) for tokenization, sentence splitting, POS tagging, and syntactic parsing
- exclude dialogue, like Pakhomov et al. (2011) and Garrard et al. (2005), which can be done reliably since e-books use typographic quotation marks
Complexity Measures (1)

Vocabulary richness:

- vocabulary size $V$ and type-token ratio $V/N$ (TTR)
- Yule’s $\kappa$ (probability of sampling the same word twice in a row)
- Honoré H ($\sim$ proportion $V_1/V$ of hapax legomena = Zipf slope $1/a$)
Complexity Measures (1)

Vocabulary richness:

- vocabulary size $V$ and type-token ratio $V/N$ (TTR)
- Yule’s $\kappa$ (probability of sampling the same word twice in a row)
- Honoré H ($\sim$ proportion $V_1/V$ of hapax legomena = Zipf slope $1/a$)

Readability & morphology:

- proportions of different word classes, noun-verb ratio
- proportion of words learnt after the age of 9 years, based on Kuperman, Stadthagen-Gonzalez, and Brysbaert (2012)
Complexity Measures (1)

Vocabulary richness:
- vocabulary size $V$ and type-token ratio $V/N$ (TTR)
- Yule’s $\kappa$ (probability of sampling the same word twice in a row)
- Honoré H ($\sim$ proportion $V_1/V$ of hapax legomena = Zipf slope $1/a$)

Readability & morphology:
- proportions of different word classes, noun-verb ratio
- proportion of words learnt after the age of 9 years, based on Kuperman, Stadthagen-Gonzalez, and Brysbaert (2012)

Syntactic domain:
- average number of words or clauses per sentence
- Yngve and Frazier depth of parse tree, which put a higher weight on left-branching sentences
Complexity Measures (2)

N-gram models (Wankerl, Nöth, and Evert 2016)

- statistical language models based on n-gram probabilities
- *perplexity* of language model determines how well part of text can be predicted from other parts
- hence gives good indication of lexical and syntactic diversity

Advantages

- language-independent
- no expensive linguistic preprocessing
Results (1)

- marked decline at the end for noun-verb ratio
- but also fluctuations among the earlier novels → decline significant?
- Yule’s $\kappa$ and many other measures don’t show expected trend
Results (1)

- marked decline at the end for noun-verb ratio and Honoré $H$
- but also fluctuations among the earlier novels $\rightarrow$ decline significant?
- Yule’s $\kappa$ and many other measures don’t show expected trend
Methodology (1)

Traditional approaches to significance testing

A assume that each text is a random sample of tokens
- binomial test with Gaussian approximation for measures based on counts
- LNRE models (Baayen 2001) for measures of vocabulary richness

B or treat each text as a single item
- Student’s $t$-test for independent samples + confidence intervals
- linear regression on time of publication (Le et al. 2011)
- non-parametric tests if normality assumption is questionable

Not valid here because

A severely underestimates sampling variation $\rightarrow$ inflated significance
B cannot be used to test significance of a single text ($n_1 = 1$ vs. $n_2 = 18$)
Methodology (1)

Traditional approaches to significance testing

A assume that each text is a random sample of tokens
   • binomial test with Gaussian approximation for measures based on counts
   • LNRE models (Baayen 2001) for measures of vocabulary richness

B or treat each text as a single item
   • Student’s t-test for independent samples + confidence intervals
   • linear regression on time of publication (Le et al. 2011)
   • non-parametric tests if normality assumption is questionable

Not valid here because

A severely underestimates sampling variation → inflated significance
B cannot be used to test significance of a single text ($n_1 = 1$ vs. $n_2 = 18$)
Methodology (2)

Our goal is to estimate sampling variation from a single text.
Methodology (2)

Our goal is to estimate sampling variation from a single text.

Key idea: **bootstrap resampling**

- cut text into equally-sized slices (*folds*)
- resample with replacement from folds
- estimate sampling variation of desired measure from bootstrap samples
- **problem**: works for counts and averages, but type-token measures are biased (bootstrapping underestimates vocabulary size $V, V_1, \ldots$)

Key idea: **cross-validation**

- compute measure for each fold, then macro-average over all folds
- estimate sampling variation of average from variability between folds
- accounts for non-randomness within text, but not e.g. stylistic differences
- still much more realistic than random sample of tokens
Methodology (2)

Our goal is to estimate sampling variation from a single text.

Key idea: **bootstrap resampling**

- cut text into equally-sized slices (folds)
- resample with replacement from folds
- estimate sampling variation of desired measure from bootstrap samples
- problem: works for counts and averages, but type-token measures are biased (bootstrapping underestimates vocabulary size $V, V_1, \ldots$)

Key idea: **cross-validation**

- compute measure for each fold, then macro-average over all folds
- estimate sampling variation of average from variability between folds
- accounts for non-randomness within text, but not e.g. stylistic differences
  - still much more realistic than random sample of tokens
Methodology (3)

As a first step

- partition each novel into folds of 10,000 consecutive tokens
- discard leftover tokens
- results in $k \geq 6$ folds for each novel
Methodology (3)

As a first step
- partition each novel into folds of 10,000 consecutive tokens
- discard leftover tokens
- results in \( k \geq 6 \) folds for each novel

Then
- evaluate complexity measure of interest on each fold

\[ y_1, \ldots, y_k \]

- compute macro-average as overall measure for the entire text

\[ \bar{y} = \frac{y_1 + \cdots + y_k}{k} \]

- instead of value \( x \) obtained by evaluating measure on full text
Methodology (4)

Significance testing procedure:

- standard deviation $\sigma$ of individual folds estimated from data

\[
\sigma^2 \approx s^2 = \frac{1}{k-1} \sum_{i=1}^{k} (y_i - \bar{y})^2
\]
Significance testing procedure:

- standard deviation $\sigma$ of individual folds estimated from data

\[ \sigma^2 \approx s^2 = \frac{1}{k-1} \sum_{i=1}^{k} (y_i - \bar{y})^2 \]

- standard deviation of macro average can be computed as

\[ \sigma_{\bar{y}} = \frac{\sigma}{\sqrt{k}} \approx \frac{s}{\sqrt{k}} \]
Significance testing procedure:

- standard deviation $\sigma$ of individual folds estimated from data

$$\sigma^2 \approx s^2 = \frac{1}{k-1} \sum_{i=1}^{k} (y_i - \bar{y})^2$$

- standard deviation of macro average can be computed as

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{k}} \approx \frac{s}{\sqrt{k}}$$

- asymptotic 95% confidence intervals are then given by

$$\bar{y} \pm 1.96 \cdot \sigma_{\bar{y}}$$
Methodology (4)

Significance testing procedure:

- standard deviation $\sigma$ of individual folds estimated from data

\[
\sigma^2 \approx s^2 = \frac{1}{k-1} \sum_{i=1}^{k} (y_i - \bar{y})^2
\]

- standard deviation of macro average can be computed as

\[
\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{k}} \approx \frac{s}{\sqrt{k}}
\]

- asymptotic 95% confidence intervals are then given by

\[
\bar{y} \pm 1.96 \cdot \sigma_{\bar{y}}
\]

- comparison of samples with Student’s $t$-test, based on pooled cross-validation folds (feasible even for $n_1 = 1$)
Methodology (4)

But is $\bar{y}$ a valid substitute for the full-text measure $x$?
Methodology (4)

But is \( \bar{y} \) a valid substitute for the full-text measure \( x \)?

- for all measures based on frequency counts (proportions of word categories) or averages (sentence length)

\[
\bar{y} \approx x
\]

(except for marco- vs. micro-averaging)
Methodology (4)

But is $\bar{y}$ a valid substitute for the full-text measure $x$?

- for all measures based on frequency counts (proportions of word categories) or averages (sentence length)
  
  $\bar{y} \approx x$

  (except for marco- vs. micro-averaging)

- n-gram perplexity is computed by cross-validation anyway
Methodology (4)

But is $\bar{y}$ a valid substitute for the full-text measure $x$?

- for all measures based on frequency counts (proportions of word categories) or averages (sentence length) 
  $$\bar{y} \approx x$$

(except for marco- vs. micro-averaging)

- n-gram perplexity is computed by cross-validation anyway
- type-token statistics might show a substantial difference 
  $$\bar{y} < x \text{ or } \bar{y} > x$$

but only if $x$ systematically depends on text size

- $\bar{y}$ then allows for meaningful comparison of different text lengths
- e.g. cross-validated type-token ratio = standardised TTR
Theoretical Validation

We can draw random samples from the Zipf-Mandelbrot law

$$\pi_i = \frac{C}{(i + b)^a}$$

to explore the appropriateness of $\bar{y}$ as a measure of vocabulary richness in the idealised situation underlying traditional binomial tests.

- Is $\bar{y}$ consistent with intuitive expectations for different fold sizes?
  - assumption: vocabulary richness determined by Zipf slope (larger values $a > 1$ result in smaller vocabulary and proportion of hapax legomena)
- How do cross-validated confidence intervals for $\bar{y}$ compare to binomial full-text confidence intervals for $x$?
Measures of vocabulary richness depend on fold size

Frequency spectrum at 100k tokens

Yule's $K$

Honoré $H$
Measures of vocabulary richness depend on fold size

- Frequency spectrum at 100k tokens
- Type–token ratio (TTR)
- Yule’s K
- Honoré H

Evert / Wankerl / Nöth (CL 2017) | FAU | Reliable measures of syntactic and lexical complexity: The case of Iris Murdoch | 25 July 2017
Measures of vocabulary richness depend on fold size

Frequency spectrum at 100k tokens

Yule’s K

Honoré H

type–token ratio (TTR)
Cross-validated vs. full-text measures

- two Zipf-Mandelbrot populations from group A
- fixed fold size of $N_0 = 1000$ tokens (confidence intervals too small for 10k)
- length of full text differs ($k$ folds $\rightarrow k \cdot N_0$ tokens)
Cross-validated vs. full-text measures

- two Zipf-Mandelbrot populations from group A
- fixed fold size of $N_0 = 1000$ tokens (confidence intervals too small for 10k)
- length of full text differs ($k$ folds $\rightarrow k \cdot N_0$ tokens)
Results (2)

- confidence intervals show that most fluctuations among earlier novels can be explained by sampling variation
- *Jackson’s Dilemma* shows significantly lower complexity according to Honoré *H*, but confidence intervals for noun-verb ratio overlap
Results (3)

Hypothesis test for equal complexity

- sample 1: *Jackson’s Dilemma* ($k_1 = 6$ folds)
- sample 2: first 17 novels ($k_2 = 151$ folds in total)
  - penultimate novel (*The Green Knight*) excluded because of unclear status
Results (3)

Hypothesis test for equal complexity

- sample 1: *Jackson’s Dilemma* ($k_1 = 6$ folds)
- sample 2: first 17 novels ($k_2 = 151$ folds in total)
  - penultimate novel (*The Green Knight*) excluded because of unclear status
- Honoré $H$: $t = -6.100$, df=5.52, $p = .0012^{**}$
  - 95% confidence interval: decline in complexity by at least 155 points (6.1%)
  - ANOVA finds marginally significant fluctuations in sample 2 ($p = .0231$)
Results (3)

Hypothesis test for equal complexity

- sample 1: *Jackson’s Dilemma* ($k_1 = 6$ folds)
- sample 2: first 17 novels ($k_2 = 151$ folds in total)
  - penultimate novel (*The Green Knight*) excluded because of unclear status
- Honoré $H$: $t = -6.100$, df=5.52, $p = .0012^{**}$
  - 95% confidence interval: decline in complexity by at least 155 points (6.1%)
  - ANOVA finds marginally significant fluctuations in sample 2 ($p = .0231$)
- noun-verb ratio: $t = -2.414$, df=5.33, $p = .0574$ n.s.
  - fluctuations in sample 1 consistent with sampling variation
Results (3)

Hypothesis test for equal complexity

- sample 1: *Jackson’s Dilemma* ($k_1 = 6$ folds)
- sample 2: first 17 novels ($k_2 = 151$ folds in total)
  - penultimate novel (*The Green Knight*) excluded because of unclear status
- Honoré $H$: $t = -6.100$, df=5.52, $p = .0012^{**}$
  - 95% confidence interval: decline in complexity by at least 155 points (6.1%)
  - ANOVA finds marginally significant fluctuations in sample 2 ($p = .0231$)
- noun-verb ratio: $t = -2.414$, df=5.33, $p = .0574$ n.s.
  - fluctuations in sample 1 consistent with sampling variation

Mixed results for other measures

- syntactic measures not significant, in accordance with Le et al. (2011)
- proportion of words acquired beyond the age of 9 shows a significant decline
- perplexity declines towards the end, but not significant
Results (3)

Hypothesis test for equal complexity

- sample 1: *Jackson’s Dilemma* ($k_1 = 6$ folds)
- sample 2: first 17 novels ($k_2 = 151$ folds in total)
  - penultimate novel (*The Green Knight*) excluded because of unclear status
- Honoré $H$: $t = -6.100$, df=5.52, $p = .0012^{**}$
  - 95% confidence interval: decline in complexity by at least 155 points (6.1%)
  - ANOVA finds marginally significant fluctuations in sample 2 ($p = .0231$)
- noun-verb ratio: $t = -2.414$, df=5.33, $p = .0574$ n.s.
  - fluctuations in sample 1 consistent with sampling variation

Mixed results for other measures

- syntactic measures not significant, in accordance with Le et al. (2011)
- proportion of words acquired beyond the age of 9 shows a significant decline
- perplexity declines towards the end, but not significant

Evidence for significant decline in *The Green Knight* (e.g. Honoré $H$: $p = .0053$)
Conclusion

- measures of lexical and syntactic complexity important for many applications, but often lack of significance testing (or inappropriate tests)
- our claim: can estimate sampling variation by bootstrapping/cross-validation
  - confidence intervals and significance tests based on normal approximation
  - our data show now evidence against normality
- assumption: texts are random samples of folds (from a single population)
  - accounts for most of the variability between earlier novels
- Honoré $H$ shows significant decline, even in the penultimate novel

Thank you!
References


