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

Stefan Evert, Sebastian Wankerl and Elmar Nöth
(Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany)

1 Introduction

Quantitative measures of the syntactic and lexical complexity of natural language text – such as type-token ratio (TTR), Yule's *K* (1944) or Yngve depth (Yngve, 1960) – play a central role in stylometric analysis. They have been used to investigate stylometric differences between writers and settle questions of disputed authorship (Stamatatos, 2009), to explore the characteristics of translated texts (Volansky, Ordan, & Wintner, 2015), to identify determinants of style in scientific writing (Bergsma, Post, & Yarowsky, 2012), to study diachronic changes in grammar (Bentz, Kiela, Hill, & Buttery, 2014), to assess the readability and difficulty level of a text (Graesser, McNamara, Louwerse, & Cai, 2004; Collins-Thompson, 2014), and as a feature in the multivariate analysis of linguistic variation (Biber, 1988; Diwersy, Evert, & Neumann, 2014).

In particular, several recent studies (Garrard, Maloney, Hodges, & Patterson, 2005; Pakhomov, Chacon, Wicklund, & Gundel, 2011; Le, Lancashire, Hirst, & Jokel, 2011) attempt to detect early symptoms of dementia in the last novels written by the British author Iris Murdoch, who was diagnosed with Alzheimer's disease in 1997. These studies focus primarily on quantitative complexity measures, based on the assumption that beginning dementia reduces either the lexical or the syntactic complexity of a patient's writing. Results were inconclusive: while the first two studies obsereved a promising decline of complexity in Murdoch's last novel *Jackson's Dilemma* published in 1995,[1] Le et al. (2011) analyzed a larger sample of Murdoch's writings and found that most of the quantitative measures did not to show any clear effects. In particular, they rejected the hypothesis of a decline in syntactic complexity.

Like most work in stylometry, all three studies fail to take the sampling distributions of complexity measures into account. As a result, they are prone to over-interpreting observed differences that may well be explained by random variation. Only Le et al. (2011) apply significance tests, but they test for a linear trend in complexity across the span of Murdoch's writing career, which would not be consistent with the typical development of Alzheimer's disease.

In this paper, we propose a novel methodology for the computation of reliable confidence intervals and significance tests for measures of linguistic complexity, inspired by ideas from bootstrapping and cross-validation. As an illustration, we apply the new method to the case of Iris Murdoch, showing that most of the differences observed in previous work are not signficant and can indeed be accounted for by sampling variation.

2 Case study: The writings of Iris Murdoch

Iris Murdoch was one of the most renowned British writers of the post-war era. Between 1954 and 1995 she published a total of 26 novels and several non-fiction works on philosophy. Most of her novels were well received by literary critics and are known for their sophisticated topics (Spear, 1995). However, her last novel was received "without enthusiasm" and Murdoch

[1] And, for some measures, also in the penultimate novel published in 1993.

revealed that she experienced difficulties while composing the work (Garrard et al., 2005). Since this novel was published only a few years before her diagnosis, it is plausible to assume that her writing was already affected by the beginning dementia.

In our case study, we attempt replicate the results of prior studies by looking at 19 of Murdoch's 26 novels, including the nine final ones. This provides complete coverage of Murdoch's late work, which is of particular interest for the diagnosis of dementia and spans a period of almost 20 years. Our two main research questions are: (i) To what extent are the conclusions of prior research affected or even invalidated if sampling variation is taken into account? (ii) Does the onset of Alzheimer's disease manifest in a clearly visible and significant decline of complexity (according to one or more of the quantiative measures)?

We purchased all 19 novels as e-books in epub file format, enabling us to extract the text directly without errors introduced by OCR software. We then used Stanford CoreNLP (Manning et al., 2014) for tokenization, sentence splitting, lemmatization, part-of-speech tagging, and syntactic parsing of the texts. Following Garrard et al. (2005) and Pakhomov et al. (2011), who excluded direct speech from their experiments, we flagged all such sentences in the novels. Detection of direct speech passages was greatly simplified by the epub format. However, the formatting of one of the e-books[2] did not distinguish between quotation marks and apostrophes, so it had to be excluded from some of our experiments.

We report our findings for a wide range of complexity measures on the novels. Most of these measures consider lexical and syntactic aspects of the writing. From the lexical domain, we evaluate the vocabulary size and type-token ratio, the proportions of different word classes, Yule's *K* (which gives the probability of sampling the same word twice in a row) and Honoré *H* (which measures the number of hapax legomena). In addition, we assess the approximate age-of-acquisition of the words appearing in a novel, using the word lists provided by Kuperman, Stadthagen-Gonzalez, and Brysbaert (2012) with a particular focus on the proportion of words learnt during later childhood and adolescence (beyond the age of 9 years). From the syntactic domain we evaluate simple measures like the average number of words or clauses per sentence. This is complemented by Yngve and Frazier depth, two measures which assign a higher weight to heavily left-branching sentences that are assumed to demand more working memory.

As a novel approach to measuring complexity, we also evaluate the use of statistical language models based on n-gram probabilities (Goodman, 2001). Such language models determine how well the text in one part of a novel can be predicted from the other parts of the novel. The *perplexity* of the model, a measure frequently used in speech processing, gives a good indication of the lexical and syntactic diversity of a text. This approach also has the advantages of being language-independent and not relying on complex linguistic pre-processing (Wankerl, Nöth, & Evert, 2016).

3 Bootstrapping confidence intervals and significance tests

Traditional significance tests based on a binomial sampling distribution cannot easily be translated to complexity measures, for two reasons: (i) they make the highly unrealistic assumption that a text (such as one of Murdoch's novels) is a random sample of individual words or sentences; and (ii) they only apply to measures based on frequency counts or other numeric averages (such as mean sentence length). On the other hand, treating each text as a single item is often impracticable: in our case, it would be difficult to find any significant effects in a sample of size $n = 19$, and it would be impossible to test for a significant deviation of a single text.

[2] *A Severed Head* (1961)

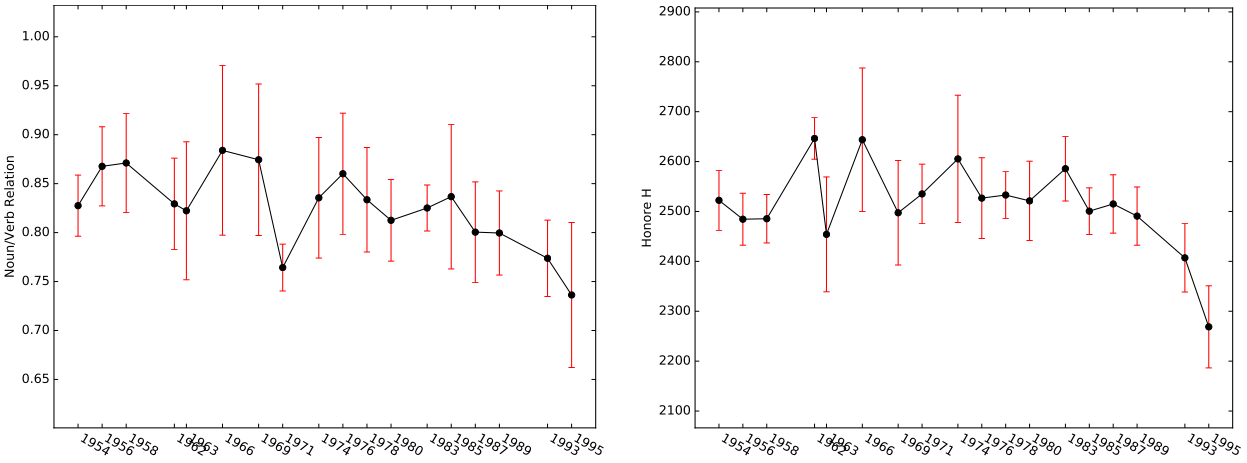

Figure 1: Bootstrapped confidence intervals for noun/verb ratio (left) and Honoré *H* (right) in 18 novels written by Iris Murdoch. The horizontal axis shows the year of publication.

Our solution combines the idea of bootstrapping (Efron, 1979) with the cross-validation procedure applied in machine learning and computational linguistics. We partition the sentences of each novel into consecutive bins of 10,000 tokens each (leading to different numbers of bins depending on text length). Leftover tokens at the end of the novel are discarded. In our case study, we obtained $n \geq 6$ bins for all 19 novels.

The complexity measure of interest is then evaluated separately on each bin, resulting in n values $y_1, \ldots, y_n$ for a given text. We compute the mean

$$\mu_y = \frac{y_1 + \ldots + y_n}{n}$$

as an overall measure for the entire text, as well as the standard deviaton σ_y across the n bins. For a measure based on frequency counts or averages, μ_y corresponds to the value that would be computed directly for the entire text. For the type-token ratio, μ_y corresponds to the standardized type-token ratio (cf. Kubát & Milička, 2013), allowing a meaningful comparison of texts of different length. The same holds for other length-dependent measures of lexical complexity.

The sampling distribution of the overall value μ_y can be determined from the standard deviation across bins. According to bootstrapping theory, an approximate 95% confidence interval for μ_y is given by

$$\mu_y \pm 1.96 \cdot \frac{\sigma_y}{\sqrt{n}}$$

The scaling factor $\sqrt{n}$ corrects for the larger variability of measurements in bins of 10,000 tokens compared to entire texts. Significance tests for different hypotheses can be constructed in a similar way, e.g. to determine whether a single text differs significantly from the remaining body of work of an author, or whether there is a significant difference between two groups of texts.

As an example, Figure 1 shows μ_y with 95% confidence intervals for two quantitative complexity measures: the ratio between nouns and verbs (left panel) and Honoré *H* (right panel). Without confidence intervals, both plots would give a similar impression of a slight decline in complexity towards the end of Murdoch's writing career. However, the statistical uncertainty inherent in the noun/verb ratio is much larger and its decline for the last two novels is not significant (as a rule of thumb, two texts are significantly different if their confidence intervals do not overlap). Honoré *H*, by contrast, is significantly lower in *Jackson's*

Dilemma than in most other novels written by the author, suggesting that it might indeed reflect a cognitive impairment.

The remaining measures yield mixed results. None of the syntactic measures shows any significant change at the end of the writing career, in accordance with the conclusions of Le et al. (2011). From the lexical domain, the proportion of words acquired beyond the age of 9 shows a significant decline in the last novel (similar to Honoré *H*). The proportions of word classes do not show any significant trend. In particular, the number of pronouns fluctuates during the entire writing career, while the number of nouns remains more stable and increases only slightly towards the end. The number of adjectives increases in the beginning, remains stable for some time and slightly decreases at the end. The type-token ratio shows a non-significant decline in the last novel. The perplexity of n-gram models remains relatively low during the first third of the author's writing career, then soars in the early seventies, and declines at the end. However, the confidence interval for the last novel is particularly wide so that no significant change can be observed.

References

Bentz, C., Kiela, D., Hill, F., & Buttery, P. (2014). Zipf's law and the grammar of languages: A quantitative study of Old and Modern English parallel texts. *Corpus Linguistics and Linguistic Theory*, *10*(2), 175–211.

Bergsma, S., Post, M., & Yarowsky, D. (2012). Stylometric analysis of scientific articles. In *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: Human language technologies (naacl-hlt 2012)* (pp. 327–337). Montréal, Canada.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Collins-Thompson, K. (2014). Computational assessment of text readability: A survey of current and future research. *International Journal of Applied Linguistics*, *165*(2), 97–135.

Diwersy, S., Evert, S., & Neumann, S. (2014). A weakly supervised multivariate approach to the study of language variation. In B. Szmrecsanyi & B. Wälchli (Eds.), *Aggregating dialectology, typology, and register analysis. linguistic variation in text and speech* (pp. 174–204). Linguae et Litterae: Publications of the School of Language and Literature, Freiburg Institute for Advanced Studies. Berlin, Boston: De Gruyter.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, *7*(1), 1–26.

Garrard, P., Maloney, L. M., Hodges, J. R., & 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.

Goodman, J. T. (2001). A bit of progress in language modelling. *Computer Speech and Language*, *15*(4), 403–434.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 193–202.

Kubát, M. & Milička, J. (2013). Vocabulary richness measure in genres. *Journal of Quantitative Linguistics*, *20*(4), 339–349.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods*, *44*(4), 978–990.

Le, X., Lancashire, I., Hirst, G., & 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*, *26*(4), 435–461.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for computational linguistics (acl) system demonstrations* (pp. 55–60).

Pakhomov, S., Chacon, D., Wicklund, M., & Gundel, J. (2011). Computerized assessment of syntactic complexity in Alzheimer's disease: A case study of Iris Murdoch's writing. *Behavior Research Methods*, *43*(1), 136–144.

Spear, H. D. (1995). *Iris murdoch*. Modern Novelists. Palgrave Macmillan.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society For Information Science and Technology*, *60*(3), 538–556.

Volansky, V., Ordan, N., & Wintner, S. (2015). On the features of translationese. *Literary and Linguistic Computing*, *30*(1), 98–118.

Wankerl, S., Nöth, E., & Evert, S. (2016). An analysis of perplexity to reveal the effects of Alzheimer's disease on language. In *Itg-fachbericht 267: Speech communication* (pp. 254–259). Paderborn, Germany.

Yngve, V. H. (1960). A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society*, *104*(5), 444–466.

Yule, G. U. (1944). *The statistical study of literary vocabulary*. Cambridge: Cambridge University Press.