Jagiellonian University
Institute of English Studies

Jeremi K. Ochab

Computational stylistics and authorship attribution: what it measures and why it works

Thesis presented in part fulfilment
of the requirements for the degree
of Master of Arts at
the Jagiellonian University of Kraków
written under the supervision of
dr Jan Rybicki

Kraków 2014

Uniwersytet Jagielloński
Instytut Filologii Angielskiej

Jeremi K. Ochab

Stylometria i atrybucja autorska: co mierzy i dlaczego działa.

Praca napisana pod kierunkiem
dra Jana Rybickiego

Kraków 2014

to

.

.

.

myself

.

.

.

so that you do not forget

about your dreams

dear old friend

Table of contents

Acknowledgements

First and foremost, I do thank all the Anglicists with whom I had the privilege to study and, in truth, share a third of my life – their names are legion; I try to remember You all.

Second, but still foremost, I would like to thank Mrs Kamila Dudzińska, THE secretary, for all these long years.

Third, but fairly above all, I express my gratitude to the Abbott and Costello of stylometry: dr hab. Maciej Eder and my supervisor dr Jan Rybicki, for their guidance, precious remarks, and believing in and bearing with me in general. I cry of fright at the mere thought of what this thesis would have to look like, should it not be for them.

Special thanks go to Michał Strojek (for the *Discworld* novels), Jakub Szpak, Anna Filipek, Paulina Wątor and others who shared with me their corpora – it's a pity I haven't been able to use them all at this time – and a translatorial *catfish*. Many thanks to José Calvo, who provided me with a lavish selection of inspiring quotations from Mary Shelley. I am grateful to Prof. Dr. Heike Zinsmeister for providing me with the relevant information on universal tag-sets, and to Aaron Plasek for inspiring discussions and some additional references.

Lastly, I owe greatly to prof. dr hab. Zdzisław Burda, a mentor and the advisor of my doctoral thesis in physics, who turned a blind eye to my unworthy linguistic inclinations.

I perceived the necessity of becoming acquainted with more languages than that of my native country. Now I am twenty-eight and am in reality more illiterate than many schoolboys of fifteen.

Mary Shelley
Frankenstein or The Modern Prometheus, 1818

Introduction

No matter how much I would try, my introduction to the history and scope of the field, to which this thesis is intended to contribute, could not possibly be more concise than what has already been written:

> computational text analysis has been used to study problems related to style and authorship for nearly sixty years. As the field has matured, it has incorporated elements of some of the most advanced forms of technical endeavor, including natural language processing, statistical computing, corpus linguistics, and artificial intelligence. It is easily the most quantitative approach to the study of literature, the oldest form of digital literary study, and, in the opinion of many, the most scientific form of literary investigation. (Ramsay 2008)

Indeed, in this thesis I used natural language processing (NLP) tools, statistical computing, and machine learning techniques, but all these might be perceived as a far-flung corner of the realm usually associated with "literary investigation." Although I strive mainly for methodological improvements, the ultimate motivation for the thesis extends over a whole range of hermeneutical questions that can be asked (but not necessarily answered!) with the computational stylistics tools.

Whether there is a common ground of literary criticism and computational tools, pertaining not only to their aims but also their means, can be inferred from the judgement that:

> The critic who endeavors to put forth a "reading," puts forth not the text, but a new text in which the data has been paraphrased, elaborated, selected, truncated, and transduced. … In every case, what is being read is not the "original" text, but a text transformed and transduced into an alternative vision, in which, as Wittgenstein put it, we "see an aspect" that further enables discussion and debate. (Ramsay 2008)

The quote elucidates just as much the nature of computational stylistics – it is not a computer juggling abstract numbers, but a way of reading and transforming a text, not dissimilar to the selective processing a human mind.

Ramsay (ibidem) further discusses the viability of computational, data driven methods for literary criticism, and sets some tenets of the form they should take. Whereas the general attitude towards the scientific method is encouraging, the author points to instances of when such methods fail, and is of the opinion that "for most forms of critical endeavour, however, appeals to 'the facts' prove far less useful." I would hope that such statement is not so absolute as to mean uselessness to "critical endeavour" but rather to the "endeavours of critics," which might have been spawned by lack of understanding and collaboration between communities of researchers:

> As has often been noted, quantitative analysis has not had much impact on traditional literary studies. Its practitioners bear some of the responsibility for this lack of impact because all too often

> quantitative studies fail to address problems of real literary significance, ignore the subject-specific background, or concentrate too heavily on technology or software." (Hoover 2008)

Unfortunately, I do subscribe my name to such bad practices, somewhat wilfully, since developing and testing efficient and rigorous tools is needed in order to ask questions of real significance. The reader is humbly asked to turn a blind eye to the shortcomings quoted above and present in this thesis.

Henceforth, I will not develop an argument for the cause of digital humanities, computational stylistics, etc. other than just showing my recent work in this field, the results of which will speak for themselves, and for better or worse. I do hope it is but a prelude to a more comprehensive research – in the meanwhile, what is presented is mainly case studies, which just perhaps are a tiny step in making it to the "the 'accelerated writing' paradigm [from which] will come the next generation of word processors — true text processors — that will give us advanced tools for automatic generation of text" (Winder 2008).

The thesis is structured in the following way:

- **Chapter 1:** "Stylometry" briefly discusses why quantifying text may be revealing in the first place and it frames the main foci of this work; its sections provide necessary background as to what can be measured, how it is measured here, and what happened to results from such measurements that could be relevant to translation studies.
- **Chapter 2:** "Disentangling grammar, topic, translation" presents a simple generic framework of how to measure the contribution of various factors to the authorial style, and contains sample results for grammatical and lexical factors followed by an exploratory study of whether language transfer can be computationally traced in translation.
- **Chapter 3:** "Using methods of community detection to attribute authorship" is rather technically oriented, with a short introduction to how graphs can represent data, and how a relatively new family of unsupervised clustering methods compare to other methods in the service of authorship attribution.
- Descriptions of the corpora analysed in this thesis, as well as listings of the novels they contain can be found in **Appendix A**, whereas the technicalities concerning NLP, stylometry, and clustering software in the **Appendix B**.

I was required to exchange chimeras of boundless grandeur for realities of little worth.

Mary Shelley
Frankenstein or The Modern Prometheus, 1818

1. Stylometry

Before increasingly narrowing down the discussion to the detailed means that stylometry uses, I would like to share a few thoughts, mostly those of other people's, on what stylometry is or strives to be and what are its intricacies. It seems that stylometry is an extension of modern authorship attribution studies in terms of its tools, and of stylistic analysis in terms of its aims, and so is also dubbed computational stylistics, perhaps for politeness's sake to distinguish itself from the traditional stylistics scholars. The difference, nevertheless, rather seems to be the methodologies and the areas fruitfully annexed by them, because in the end:

> Stylistic analysis is open-ended and exploratory. … Authorship studies aim at "yes or no" resolutions … Yet stylistic analysis needs finally to pass the same tests of rigor, repeatability, and impartiality as authorship analysis if it is to offer new knowledge. And the measures and techniques of authorship studies must ultimately be explained in stylistic terms if they are to command assent. (Craig 2004)

That is, phrasing it differently, it is rather the authorship attribution methods that actually utilise or are part of stylometry, because supposedly they distinguish between authors based on their style, measured in a somewhat abstract way. The reason for such a hierarchy has also been expressed by Hoover (2008):

> Authorship attribution and statistical stylistics (or stylometry), currently two of the most important areas of quantitative analysis of literature … share many basic assumptions and methods, though some techniques that are effective in distinguishing authors may have no clear interpretive value. … literary attribution is often only a first step, so that methods easily turned to stylistic or interpretive purposes tend to be favored.

Then, of course, it is a well posed question, how does or can stylometry contribute to stylistics, and whether the new methods are enough to make a qualitative jump.

> In such work, a little paradoxically, one wants to be reassured by seeing patterns already familiar from the way the texts are usually discussed, yet also to be surprised so that they seem more than a restatement of the obvious. (Craig 2004)

which means that we hope stylometry will not only provide a firm methodological foundation for stylistics, but also raise new question and "boldly go where no stylists has gone before." Debatable as these issues are, we need to engage into some more applicable methodological ponderings.

Craig (2004) provides a much broader historical and scholarly background (with a strong Anglo-Saxon perspective, one must note) in the short introduction to authorship attribution and stylometry and the relationship between the two. He mentions various premises concerning why we can expect to distinguish writers based on statistical variation of their writing or what the social and cognitive circumstances that lead to this variation are. Among others, he cites some of the reservations about the methods. In this thesis, at least partly or parenthetically I will address some of the pitfalls that he mentions:

> 1. Assuming that, if groups separate according to author, the separation is authorial.
> …
> 4. Assuming that an author cannot vary from his or her normal style when carrying out a particular assignment, or under a particular internal or external influence. (Craig 2004)

To make the first assumption valid one would need to control for all the other factors like genre, topic, etc., and only then what is left is the authorial; the fourth assumption, in turn, is crucial in the study of pastiches, translations, etc.

In his Principal Component Analysis of 25 Shakespeare's plays, on seeing some patterns Craig (ibidem) notes that "one would want to know how constant these patterns are when a play is added to the set or taken away, for instance, or if different variables are used, for instance by splitting *to* into infinitive, prepositional, and adverbial uses," which are recurring nightmares of a stylometrists, i.e., dependence of the results on the corpus and on the set of features comprising the basis (incidentally, even the exact mathematical sense of the word would be correct here) for the analysis. Many such problems are shortly exemplified in the following passage:

> Beyond these are questions of the design of the experiment. Why choose words, and why these words? Why choose plays, when one might choose *oeuvres* or periods as larger aggregations, or characters or scenes as segmentations, or some combination? How would the main lines of differentiation in a larger set of Elizabethan and Jacobean plays, going beyond Shakespeare, compare? (Craig 2004)

Chapter 2 will orbit around the first of these whys, which is irrevocably married to the questions posed by Craig a bit further:

> What, for example, is the status and nature of the axes of differences that the procedure has derived? They have a precise definition – each variable and each case has a score for each vector – but this says nothing about the stylistic context. (ibidem)

I tend to think about these fundamentals in terms of a cooking recipe metaphor: the style, as seen by the present computational methods, in not just a layered hamburger – rather, it is a dough with its ingredients combined and blended, and chemically and physically bonded. What stylometry strives and struggles to achieve requires more knowledge than just the right proportions of words to mix; like a true alchemist we need to know what reactions take place. How can you make a golem from perfectly weighted amounts of oxygen (32500 g), carbon (9250 g), hydrogen (4750 g), nitrogen (160 g), etc., or how can you make a Frankenstein's living human from eyeballs, arms, and livers, for that matter? Both physical and biological and linguistic structures are emergent (Anderson et al. 1972), and unfortunately only the higher-level ones are interpretable – or rather have meaning – for us. Even such a view can be contested:

> A central problem for [Stanley] Fish is the assumption that meaning resides within the text rather than being created as it is read. He argues that the formal features described by stylisticians are meaningless except in relation to the reader's perception of them within a reading situation. (Craig 2004)

Hence, instead of discussing further whether computational stylistics is

> at best, a powerful new line of evidence in long-contested questions of style; [or] at worst, an elaborate display of meaningless patterning, and an awkward mismatch between words and numbers and the aesthetic and the statistical (ibidem)

in the next chapters I will aspire to make the first expression a step closer to the truth and the second one leap farther.

The aims, or hopes, as to what can be evidenced by the means of quantitative approaches "are most naturally associated with questions of authorship and style, but they can also be used to investigate larger interpretive issues like plot, theme, genre, period, tone, and modality" (Hoover 2008). Also Burrows (2004) makes it explicit that the different textual levels can be traced in such an analysis:

> Differences in level of discourse (as in the wide-ranging contrasts between Latinate forms like ascend and Germanic equivalents like go up); in genre (as between the past-tense verbs of retrospective narrative and the present tense of most disquisitory writing); and in historical provenance (as in the shift from thou to you) – all these and many more are reflected in systematic relationships among the frequencies of the very common words.

Choosing subsets of word frequency lists, as Burrows did, is a way of disentangling such different factors. Indeed one can choose an optimal subset of features that generate the best authorial or other distinction.

> Since this pattern is certainly not author-driven, the question is whether other inferences can be drawn. A close study of the way the poems cluster offers rough but suggestive distinctions between predominantly monologic, dialogic, narrative, and reflective forms of rhetoric.

I would like to make these observations, though, a little less anecdotal. The problem with authorship attribution often is that the interpretation of what made two texts or groups of texts close or distant (or further, clustered together or not) is made *a posteriori*, e.g., "Charles Cotton's *A Voyage to Ireland in Burlesque*, is isolated by its narrative mode and its colloquial speech-register" (Burrows 2004) or "Why do they cross the border between verse and prose? The best explanation, *I believe*, is that Stapylton favors a dialogue of brief, deictic interchange and only rarely offers a large, poetic set-speech" (ibidem; emphasis mine). In fact, Burrows already knows such high-level characteristics like narration and register and after seeing the result he assumes these are the characteristics that made the author different. This is possible, but in principle we do not know what the precise connection between word frequencies and register or narration is, and then how the clustering is affected (well, this is known a little better). [Some of the pitfalls of authorship attribution based on Nearest Shrunken Centroids, but valid in general, are discussed at length by Fields et al. (2011).]

I seek to find this connection from the corpus data itself. How much is it the register or narration, and how much is it some other idiosyncrasies of that particular author that made him stand out. Frankly speaking – as one should not in a master's thesis, – without answering that question usually no new insights can be gained, but only previous intuition are confirmed, such as: these pieces are by the same author, these pieces are the same genre, etc. Of course, even such results are valuable when comparison is made between texts that have not been previously read (which is exactly my case in the next chapters), e.g., because "computers obviously surpass our unassisted powers in managing large textual corpora, singling out unique forms or gathering all the instances of common ones" (Burrows 2004). Such assistance is irreplaceable in performing a very large-scale analysis of hundreds books, or millions for that matter, for which one does not have human work power to obtain any intuition.

To cherish even a little more hope in computational methods, if the algorithms of authorship attribution are not 100% accurate, this may be good news, for it may point to the fact that some book of an author is indeed different in some respect (which can be observed, e.g., for Virginia Woolf). This, in turn, might confirm the facts already known, but might just as well be a new unknown.

1.1. Methods

Stylometry is about measuring style. Thus, before performing a stylometric experiment we need to decide precisely what observable quantity we want to measure. Hoover (2008) makes an introductory overview of what can or has been counted in texts:

> Almost any item, feature, or characteristic of a text that can be reliably identified, can be counted, and most of them have been counted. Decisions about what to count can be obvious, problematic, or extremely difficult, and poor initial choices can lead to wasted effort and worthless results.

Stamatatos (2009) also presents a catalogue of stylometric features and the tools needed for their measurement. I list some of them below, leaving aside the semantic and most of the syntactic ones:

- characters,
- character n-grams,
- word/clause/sentence/paragraph/text lengths,
- words,
- word n-grams,
- type-to-token ratio (i.e., the number of *different* words in a text divided by the number of their occurrences in a text),
- part-of-speech (PoS) tags.

As one might expect, the character counts are not too informative; characters have no meaning by themselves; but let us assume for a moment that there are two identical-twin authors, and one overuses the word "undoubtedly", while the other "indubitably". Then, *ceteris paribus*, the frequency of the letters *d, e, o, u* would be higher for the first one, while *a, b, i* for the second one, and that's how we could distinguish between the two. In principle, under certain circumstances one could think of inferring from such information the number of sentences, clauses, or words.

In reality, since the letters from the example contain also the information on all the other hundreds of thousands words churned out by the twin authors, the difference between the their letter frequencies would be very tiny. In fact, if there were some randomness involved in their writing – that is to say, in this or that novel they make slightly different word choices, but *on average,* in all the novels they have written, both of them use the same words – then by comparing only one novel of each, the indubitable difference could be rendered doubtful or *insignificant* (which on this occasion is also an accurate statistical term, mind you). In other words, the fluctuation, i.e., the slight random difference in the word usage between single novels could be enough to submerge the subtle 'indubitable/undoubted' difference. Some discussion of the care one has to pay to the (natural) variance in the data can be seen in Hoover (2008).

There are thus two profound problems of different nature: the first concerns interpretability of the results, their explanatory power, and their connection to information content, language structure, and linguistic theories in general (what does it mean that authors have different character counts? where did it come from? did they write on different topics or in different styles?); the second concerns the mathematical laws governing information retrieval (probability, statistics, information theory, etc.), which provide tools to tell us how certain we can be that a given conclusion is true (e.g., these books have been written by different authors) given some information (e.g., these two authors usually have such and such character frequencies).

For a second let us focus on the second problem: the statistics. This involves several issues as the size of the corpus needed (5000 word text samples for EN and PL in, see Eder

(2013) or for a different machine learning method Luyckx and Daelemans (2011), proper normalisation or standardization, bootstrapping, and all the techniques warding of the menace of insignificance. The problems of too little statistics and randomness are however heavily dependent on the mere choice of what is measured. Counting individual letters will result in much smaller errors because there are just a few of them, while they occur very frequently (about 3250 on this page, which makes it 125 counts per letter on average). Counting words is less reliable, since the total number of words in English is possibly well above a quarter of a million (and there are about 600 word tokens on this page, but as much as 320 different word types, which makes it less than 2 on average). As can be easily imagined, a fluctuation of 2 character counts per 100 is fairly small, but 1 word in 2 is probably too much to lead to any significant conclusions.

Coming back to the first problem, of which I intuitively think as flattening of information (i.e., reducing all content to just several incomprehensible variables – the numbers describing letter occurrences), it is a chasm which disconnects the numerical result from a humanly interpretable linguistic content. (As a side note, this is also a chasm that may seem to separate literature and language scholars from literature and language researchers who utilise computation.) This issue does pose some threats, but is not vital for authorship attribution, where one simply wants to know who wrote what. If the question is not whether two books or authors are related, but how and why they are or are not, such simple, flattened, numerical characteristics do not suffice. As Heuser and Le-Khac (2012) remark, "the greatest challenge of developing digital humanities methods may not be how to cull data from humanistic objects, but how to analyse that data in meaningfully interpretable ways." Providing a brick for the bridge over this chasm is one of the main aims of this master's dissertation.

Just as inferring from letter counts, hypothetically, the number of sentences or words, from the frequencies of words one could attempt at inferring the number of sentences or the most probable statistics of clause types. Similarly, word n-grams ought to have inscribed some information on phrases, collocations, relationships of prepositions and verbs, some word-punctuation interplay, etc.

Again, to show how this may work with the proviso that

> Although letter n-grams lack any transparent relationship to the meaning or style of a text, and are unlikely to be attractive to researchers who are interested in broader literary questions, word n-grams are likely to become increasingly popular because they may both improve accuracy and allow the critic to focus on meaningful word groups. (Hoover 2008)

let us resort to an example of character n-grams. They already do contain some morphological information (e.g., 3-grams could detect an increased number of gerunds 'i-n-g', but such *grammatical* information is of course blended with other, e.g., both "boring", "bring", and "bingo" would spawn among others the 'ing' 3-gram), it also contains some information on word co-occurrences (when spaces are included as characters); 3-grams would also contain 1-, 2-, and 3-letter words, which includes many prepositions, pronouns, determiners and other function words. This shows that some residual, mixed information can be mined in such simple features as character n-gram frequencies.

The last possible choices of measurable quantities might involve conjoint frequency lists of n-grams of different sizes, or both character and words n-grams, etc. It is also possible to include or exclude only certain positions of such frequency lists, e.g., one could base the comparison of two authors only on the verbs, or perhaps on all the words but the ones from the maritime domain. The choice depends on the particular research question that one asks and on the data available. Additionally, as can be deduced from the

paragraphs containing comments on statistics, there is a certain counterbalance between the meaningful information content that can be measured in a text and the statistical uncertainty of that measurement. For this reason, taking too long n-grams may introduce too much *noise* (another term for a random fluctuation).

Eder (2011) compare which features (for English, German, Latin, and Polish texts) are most effective in authorship attribution: word uni-, bi-, tri-, and tetra-grams, letter bi- up to hexa-grams, a combination of words and word bi-grams, and a combination of words and letter penta-grams. The results that are relevant to us are for English: the far end of word frequency list (5000-10000), although the attributive success rate is high for the rest of words as well; for Polish: 100-500 most frequent words or 1000-2000 letter 5- or 6-grams.

In a little more recent work Rybicki and Eder (2011) Eder focused on words (i.e., unigrams) only; their "heat maps" of attributive success show that the wordlists' initial position and length is optimal at: 0-600 and 400-maximum, respectively, for EN novels, although it works almost as good pretty much everywhere; 350-450 and 350-750, respectively, for PL novel classics. Although as yet there are no more detailed studies, the optimal choices of feature types and lists seem to depend strongly on the language (supposedly, on the degree of its inflection or perhaps the literary tradition and strength of non-authorial signals).

1.2. Burrows's Delta

Since this thesis is not aimed at a comprehensive review of different methods, I stick to just one, which is based on what is called Delta distance between frequency lists (Burrows 2002a), which has proven to be reliable, and scores one of the best results in authorship attribution comparisons (as also shown in **Figure 14** in **Chapter 3** for the EN100 benchmark).

The scheme is as follows:

I. calculate (word) frequency lists for all books in the corpus; hence, we obtain N × M matrix

$$F = \begin{pmatrix} f_{11} & \cdots & f_{1N} \\ \vdots & \ddots & \vdots \\ f_{M1} & \cdots & f_{MN} \end{pmatrix}$$

where N is the number of books and M the number of words, whose elements f_{wb} are the frequencies of *w*-th word in *b*-th book (i.e., percentage of occurrences of *w* among all the tokens in the book), so that each column contains the information on one book, and each row contains information on a particular word. Mark, that we may restrict the set of words we want to take into account, e.g., by taking only the words that appear in all texts (see *culling* in **Appendix B.3**) or only the first 100 most frequent words in the corpus, or perhaps only the verbs, and so on. Other less straightforward ways of selecting words of moderate and low-frequencies are, e.g., Burrows's Zeta and Iota (Burrows 2007). Picking up just a few words is also possible, but (at the very least) with the proviso that "while useful to distinguish between one particular pair of authors, may be irrelevant when comparing another pair of authors" (Luyckx and Daelemans 2011).

II. for each word *w* calculate the mean μ_w and standard deviation σ_w of frequencies for the whole corpus

$$\mu_w = \frac{1}{N}\sum_{b=1}^{N} f_{wb}$$

$$\sigma_w = \sqrt{\frac{1}{N-1}\sum_{b=1}^{N}(f_{wb} - \mu_w)^2}$$

III. standardize frequencies f_{wb}, i.e., replace them with their z-scores (note that the z-scores are strongly corpus dependent, since the mean and standard deviation depend on the corpus that we have chosen):

$$z_{wb} = \frac{f_{wb} - \mu_w}{\sigma_w}.$$

In short, such numbers tell how many standard deviations from the corpus average does a frequency of a given word in a given book lie; or, phrased differently, how abnormally does the word behave in a given book as compared to the whole corpus. If, say, $z_{wb} > 1$, the word *w* is suspiciously more frequent in a book *b* than in the rest of the corpus, while if $z_{wb} < -1$, *w* is suspiciously less frequent than normally (in the corpus).

IV. finally, calculate *distances* between the texts in the corpus (one can also calculate distance to a model text outside corpus, e.g., a benchmark text of a given author):

$$\delta_{ba} = \frac{1}{M}\sum_{w=1}^{M}|z_{wb} - z_{wa}|,$$

which is a distance between books *a* and *b*. All such distances can once again be stored in the form of a symmetric matrix:

$$\Delta = \begin{pmatrix} \delta_{11} & \cdots & \delta_{1N} \\ \vdots & \ddots & \vdots \\ \delta_{N1} & \cdots & \delta_{NN} \end{pmatrix}.$$

Such matrix can be treated as input data for some clustering algorithms as described in **Chapter 3.1**.

Some extensions of the Burrows Delta can be found in Hoover (2004a), Hoover (2004b), Eder et al. (2013), Argamon (2008), but are not discussed here. Limitations of the Delta, some optimal parameters for the most frequent words (MFW), and improvements with a cosine distance can be found in Smith and Aldridge (2011), while some criticism in Vickers (2011).

1.3. Stylometry of translation

An overview of some most notable results in quantitative literary studies (as of 2008) is presented in Hoover (2008), and one cannot comprehensively summarize all of them without writing a whole long review paper – the wealth of methods and the range of texts studied by authorship attribution and stylometry scholars is becoming overwhelming. However, there still is a subset of problems of a subtler nature:

> Little or no attention has been paid so far to the possibility of describing the 'style' of a translator or group of translators in terms of what might be distinctive about the language they produce … not in the traditional sense of whether the style of a given author is adequately conveyed in the relevant translation but in terms of whether individual literary translators can be shown to use distinctive styles of their own. (Baker 2000)

Indeed, translation gives rise to difficult stylometric obstacles, because one cannot in principle know

a) if or how a style of a text produced in one language is systematically transferred into another language in the process,
b) how the translator influences or contaminates the style of the target text with his or her own linguistic habits, predilections, or conscious choices,
c) what happens when collaborative translation takes place (more than one translator or a translator collaborating with an editor).

For instance, the second issue might result in works by different authors appear more alike – i.e., more than in the source language – when translated by the same translator or perhaps by translators having the same background (e.g., born or working in the same place or time). Still, the works by Rybicki (2012, 2013) show that in a corpus of translated texts it is the authorial rather than the translatorial signal that is visible by the virtue of Delta method; it is, as well, rather a volume of a book, rather than a translator that is stronger. As has been shown for PL↔EN and EN↔FR translations, only in a corpus written by the same author can dendrograms (i.e., graphs alike **Figure 3**) reach the translation level.

I will summarize here two selected examples [for more, consult Hoover (2008)] of what results can be obtained thanks to stylometry:

> 1) In one of his seminal papers, Burrows (2002b) examined 15 English translations of Juvenal's *Tenth Satire* (from Henry Vaughan, 1646, to Peter Green, 1967; 4 of them in prose, the rest in verse). Four of the translators were also poets (John Dryden, Samuel Johnson, Thomas Shadwell, and Henry Vaughan), thus allowing to attribute authorship to their translations by comparison with their own works and the works of other poets of the English Restoration period (25 poets in total). Using his Delta procedure, Burrows was able to show that the stylistic signature of some authors (e.g., Dryden) disappears, when they translate. Such an observation makes it impracticable to attribute authorship of a translation by way of comparison with a translator's own pieces, or at least there might be a bit of a lottery involved.
>
> Nevertheless, the method deals well with an inverse question: knowing the author, which of the translations was most probably made by him or her? Delta can also show which of the translators best concealed their own style and which adhered to it – thus, possibly disregarding the style of the original. In fact, by producing average statistics of all the translations Johnson lies furthest from the mean, which might indicate that his work was least faithful to the original as well. Such large numerical deviations from the mean word frequency

distribution also show a way to comment on some particular usage of words by the translator (e.g., by close reading one might then understand why there are so few articles or personal pronouns in the Johnson's text).

2) The paper by Rybicki and Heydel (2013) is an example of a successful translatorial (in analogy to authorial) attribution, especially with respect to the point c) above. The case study concerned Virginia Woolf's *Night and Day* (1919) translated into Polish by Anna Kołyszko, who died in 2009 leaving the work only partly finished (some chapters finished plus a draft of some further part), which resulted in another translator, Magda Heydel, taking over: editing the drafts, translating what was left, and editing the whole. Both translators that could boast notable translatorial achievements (e.g., Heydel's translations of: Woolf's *Jacob's Room*, *A Moment's Liberty* and *Between the Acts*, Graham Swift's *The Light of Day* and Conrad's *Heart of Darkness*, and Kołyszko's translations of: McCarthy's *Child of God*, Miller's *Tropic of Capricorn*, Roth's *Portnoy's Complaint*, Rushdie's *Midnight's Children*). Additionally, Heydel, being a Woolf scholar, with a number of strategic stylistic characteristics of Woolf's text she wanted to retain, did a thorough editing, modifying especially the point of view and perspective of the narration.

The stylometric methods (once again, Burrows's Delta followed by a cluster analysis, which was bootstrapped to obtain a consensus tree diagram), however, were able to clearly distinguish which chapters were translated by Kołyszko, including the one that was only a draft, and which by Heydel. This was also one of the rare occasions that the author – well, in this case the translator – could be consulted if the attribution had been correct. How delicate such a stylometric analysis can be was further illustrated by what happens when the two parts of the book are immersed in a larger corpus of texts translated from EN to PL: the translatorial signal then became submerged by the stronger authorial one.

Of course, as stressed in Rybicki (2010), any computational stylistics apprentice should be reminded – myself being in the ranks – that the traditional methods like "biography, history, graphology, traditional stylistics" ought to be used first, and only when they fail the stylometry reserves are to be thrown into the battle. Some initial information and a hypothesis on authorship is always needed. The traditional methods notwithstanding, I am of the opinion that stylometry can also be used for a large-scale pre-processing stage, e.g., to separate the uninteresting from the suspicious texts, and only then trying the more effort intensive traditional tasks. A case study that can be referred to in support of this view is the paper by Rybicki (2010), in which he scrutinizes the corpus containing translations (by 4 different authors and 2 different languages – unless some relay translation took place) and the translator's own works (both ethnographic studies and memoires). The latter texts were known to be dictated to and edited by a wife of the translator in general, and the memoires of the husband in particular were discovered to be written by the wife, as the original manuscripts, letters, etc., had been found and examined. Nonetheless, the stylometric analysis was able to provide some *new* hypotheses: that some of the translations seem attributed (more heavily edited or co-authored) to the wife unlike the others. It seems that in traditional and computational methods can take turns.

I demand a creature of another sex, but as hideous as myself; […] It is true, we shall be monsters, cut off from all the world; but on that account we shall be more attached to one another.

Mary Shelley
Frankenstein or The Modern Prometheus, 1818

2. Disentangling grammar, topic, translation

It has already been a recurring topic in this thesis that

> Apart from individual style, various other factors determine variation in text, such as topic, genre, register, and domain. Ultimately, authorship attribution techniques should be sufficiently robust to discriminate between these interacting sources of variation. That said, keeping a maximum of these interfering factors constant, is a good strategy for finding reliable indicators of style. (Luyckx and Daelemans 2011)

Whereas the extraction of a text's topic or automatic methods of determining characteristics of genre, register, etc. are complex theoretical and computational problems in themselves, there already do exist well-developed natural language processing tools for part-of-speech (PoS) tagging, lemmatizing, or parsing texts. Therefore, in what follows I present an approach to discriminating more general factors, namely vocabulary and grammar, however vague it may sound, and further variation stemming from translation. The methodology, just as quoted above, is conceptually simple but hard to do in practice: keep things constant.

2.1. Methods: hybrid randomised texts

While already writing up the thesis I came across the article by Winder (2008), who develops an approach very much the same as I have taken; the aim, however, is different, or perhaps it is just the reverse of the same coin: Winder's efforts are directed towards computer-aided text production, "accelerated writing" as he calls it, while the direction I took was computational methods of unraveling the elements of style of existing texts.

Although our goals face each other, the ways of achieving them became surprisingly convergent. Winder's primary interest is text transmutation, e.g.,

> chimeric poetry which combines the style of two poets. One example is the Rimbaude-laire, a set of poems built from templates extracted from Rimbeau's poems and lexical items from Baudelaire's (or vice versa). (ibidem)

Hence, whenever I fail to state my point clearly, the reader's understanding of these issues will surely benefit from reading the Winder's text.

While I have not reached as far as Winder, the map he outlines is akin to my view of how to proceed: **"**automatic techniques depend on the formally described topographies of the source and target texts. Beyond characters, more important linguistic topographies are syntax, lexical meaning, and narrative" (ibidem).

The precise way of random computer generation of the hybrid (chimeric) texts with respect to grammar, as analysed in **Chapter 2.2.**, is as follows:

1) take a text (or a collection of texts) of author A and tag its parts of speech,
2) record the sequence of PoS for the text (or compute PoS distribution for A),
3) take a text (or a collection of texts) of author B and tag it too,
4) for each PoS class collect all the words that belong to it in text of B into a "bag" (the bag should either contain *all the tokens* or the distribution of types),
5) generate the new text: for each PoS tag in the sequence from 2) draw randomly a token from the bag corresponding to that PoS.

In such a way we obtain a text – non-sensical, of course – whose coarse-grained grammar has the statistical properties of author A, while it uses only the words used by author B and the word frequency lists *within each PoS class* are the same as for author B. (It should be noted that in producing the hybrid texts I do not intend to generate a readable or intelligible text. Such efforts have been made, particularly with the aim of fooling journal editors or just producing nonsense pseudo-scientific gobbledygook. As yet, I do not make use of these advances.)

The procedure is a rough version of what Winder (2008) would call "syntactic templates", which "are useful starting points for generation because they largely resolve the major systemic linguistic constraints and so set the stage for a free stylistic and referential combinatory.**"** One must complement this statement with a comment that these systemic constraints also do vary between the authors and are indicative of authors style, as will be seen in the next chapters. Winder was not concerned with this issue when generating a hybrid text.

> What is more challenging is selecting semantically relevant combinations from all the possibilities. A semantic templating system requires a different set of linguistic resources. (Winder 2008)

In the semantic domain, Winder goes on substituting stereotypically gender marked words of one sex with the other or finally substituting words with their hypernyms or synonyms or words belonging to the same semantic domain or synset – something that is yet beyond the ambit of this thesis. Winder discusses even higher level developments: those of automatic narrative generation.

Now, in what kind of problems can such a text generation procedure avail us? In the present situation, we have a tool for extracting some information from texts, e.g., authorship, but we do not have a model for a text of a given author. Or rather we have – it is the word frequency list or the like – but it is a meagre model, too simplified to account for the phenomena that we want to study. The goal now is to make the model richer, less simplified, but tractable and manipulable, still.

> In other words, a model-of is made in a consciously simplifying act of interpretation. Although this kind of model is not necessarily a physical object, the goal of simplification is to make *tractable* or *manipulable* what the modeler regards as interesting about it. (McCarty 2008; original emphasis)

What we want to trace is the specific layers of authorial style; what we want to manipulate are our null hypotheses of what contributes to the style. Thus, the model used in the procedure 1)-5) includes not just the word frequencies, but a conglomerate of words and their PoS tags, allowing to *manipulate the word frequencies indirectly* by manipulating the distributions of speech categories. And following further the line of McCarty (2008)

thought, "the exact correspondence between model and object … may be violated deliberately in order to study the consequences." That is precisely what the hybrid text generation does. It computes a model of a text by author A by following its certain statistical properties and then it violates some of these properties (the distributions of tokens in PoS bags) in order to check how much a new hypothetical text differs from the real one.

2.2. Results: Hybrid computer-generated texts

Since the jump from distance table to the clustering (extracting the authorial groups) is far from trivial, there is no way of telling how much the distance has to change in order for a fake text to be misattributed to another author. That is why at first I study merely the change in distance between two texts (which is just one entry in the table of Delta distances, which is fed into the clustering algorithm), and only then do I show how it affects grouping texts.

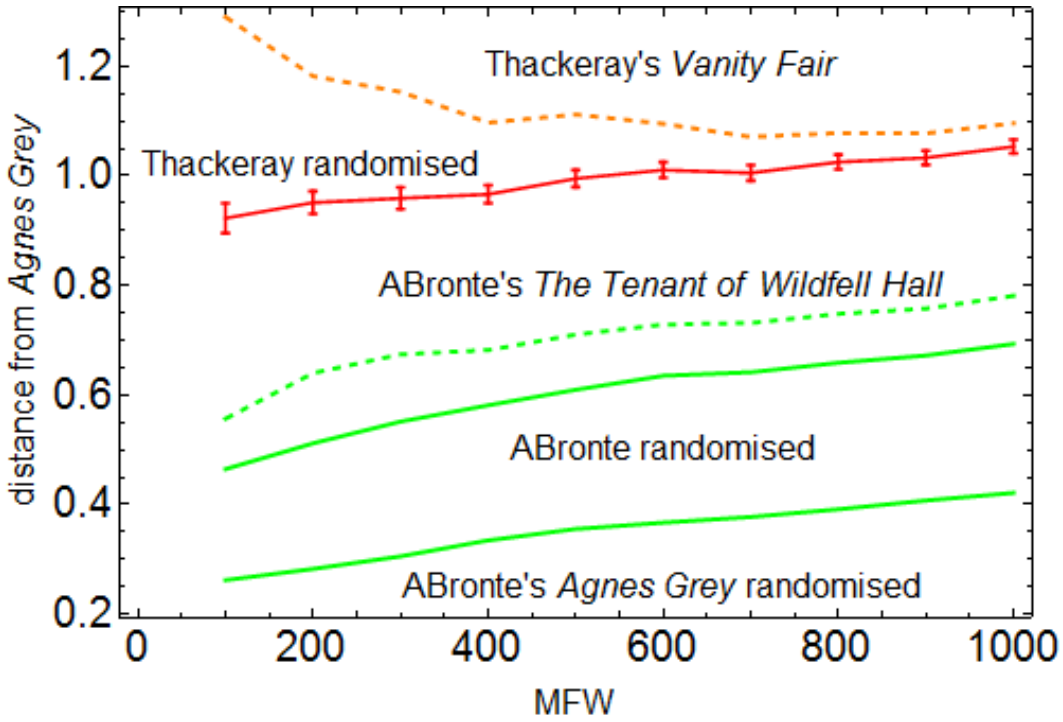

Figure 1 Burrows's Delta distance from *Agnes Grey* depending on the number of most frequent words (MFW) taken in to account. The dashed curves correspond to real novels, and the continuous lines to the hybrid computer-generated texts, respectively.

As shown in **Figure 1**, the distance between *Agnes Grey* and itself is 0, but already a distance to another novel of the same author can takes values between 0.6-0.8. The reference point for the hybrid randomization is a text that uses the exact distribution of tags from *Agnes Grey* and the concrete words are drawn from the catalogue of words for each given tag, where the vocabulary distribution reflects the distribution extracted from the book: the distance is far from zero, but it is very hard to find a pair of novels in an authorial collection that would be as close. Another continues green line represents another hybrid randomised text, where the distributions were taken from both Anne Brontë's books present in the corpus.

Next, we see that the distance between Thackeray's *Vanity Fair* and Brontë's *Agnes Grey* is much higher; especially in the regime where synsemantic words constitute majority (MFW = 100-200). Subjecting Thackeray to the same distribution of tags, however, is enough to produce a significant drop. It can already be seen that the lower number of MFW one takes into account, the more significant the distribution of tags is. This appears to be connected to the fact that the percentage of function words is much greater for the very first MFW, while for higher number of MFW the content words take over.

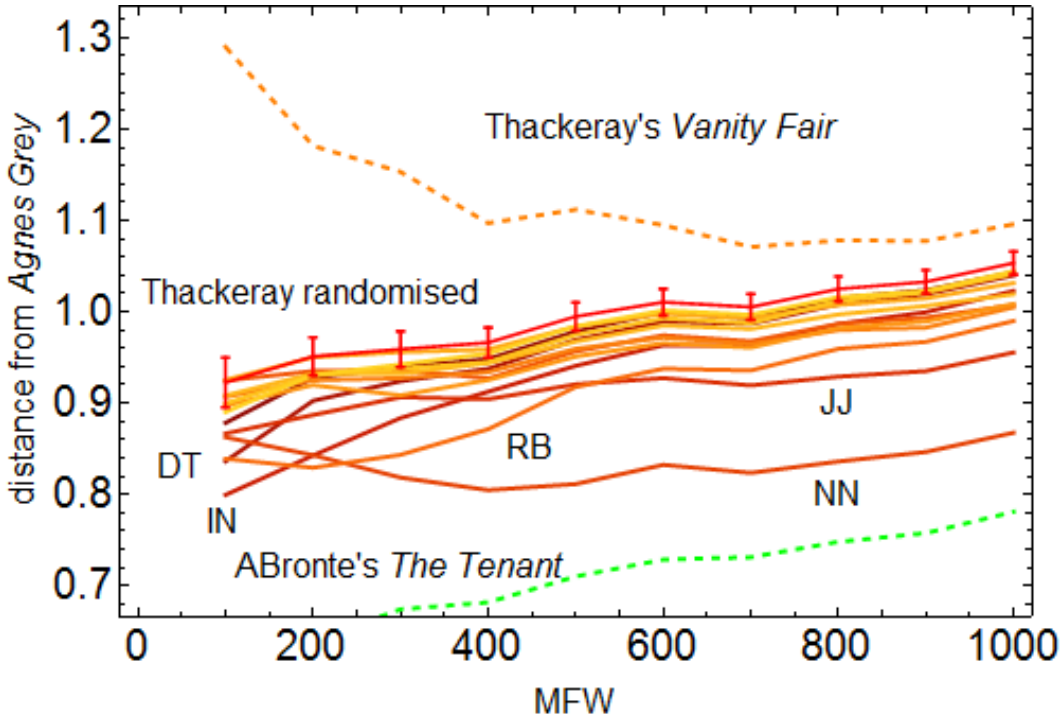

Figure 2 Burrows's Delta distance depending on the number of most frequent words (MFW), a magnification of Figure 1. Each of the orange and brownish continuous curves correspond to a randomized Thackeray's text, in which the distribution of words from a given PoS category was pasted from *Agnes Grey*. The most important PoS tags are: determiners (DT), subordinating conjunctions and prepositions (IN), nouns (NN), adjectives (JJ), and adverbs (RB).

This presupposition is confirmed in **Figure 2**, in which the determiners, subordinating conjunctions, and prepositions are the parts of speech that strongly account for the similarity between texts in the regime of 100-300 MFW, while substitution of nouns, adjectives, adverbs makes the texts more similar for greater number of MFW taken into account. (That verbs do not seem to play such an important role is probably due to the fact they are divided into several subclasses.) It can be conjectured that since the PoS categories containing synsemantic words are much smaller, it is only their distribution that distinguishes the authors (because the authors usually use quite the same range of these words), while for the open class of lexical words the substitution which made the curves to lower so much in **Figure 2** involved to much greater extent the topic or generally the content of the book. The cumulative effect of substituting all the function words or all the content words requires recalculating all the distance tables and has not been done here.

The effect of the distance change of individual PoS on authorship attribution, however, is checked within the small collection of British fiction, and can best be visualised by dendrograms below, see **Figure 3-Figure 5**. Although faking the PoS tag

distribution of A. Brontë in the hybrid text that used Thackeray's vocabulary is not enough to fool the clustering algorithms, one can already see that such a change was able to move the text to the fringes of Thackeray's authorial group, when only the first 100 most frequent words were used in the analysis. This shows that such a coarse-grained information already contains traces of authorial style that we would like to pinpoint.

The additional decrease of distance between the texts brought by the substitution of word frequency distributions of a particular part of speech already does fool the algorithm, see **Figure 4**. The IN and NN are in truth the most numerous word classes in English (around 11-12% of tokens each), as shown later, e.g., in **Figure 7**, which I assume is partly responsible for this little success. This misattribution does not take place when a consensus tree over 100-1000 MFW is drawn, but as could be expected from **Figure 2**, in which the NN line keeps low for the whole range of MFW, indeed substitution of just the nouns makes the hybrid text least similar to the rest of Thackeray's works.

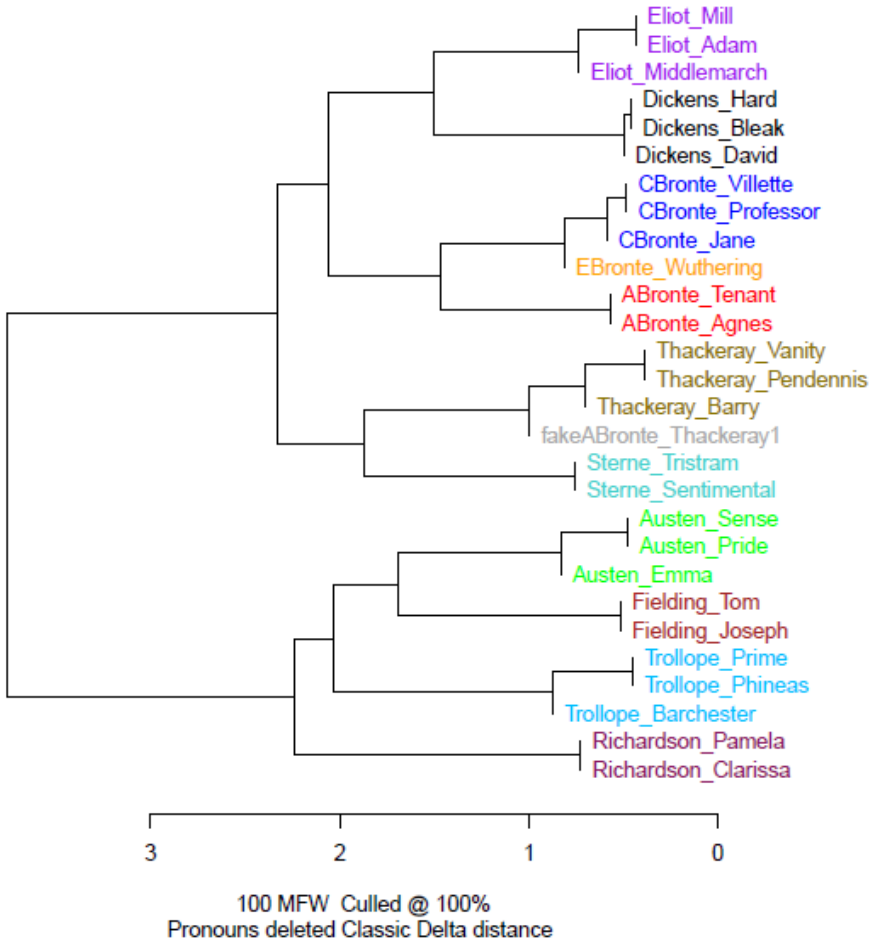

Figure 3 Dendrogram of classic British novels (see Appendix A.1) together with the computer-generated hybrid text based on Thackeray's vocabulary and A. Brontë's PoS-tag distribution. The hybrid text (labeled fakeABrontë_Thackerey1) is apparently least Thackeray-ish. (Note that there are only 100 MFW taken into account.)

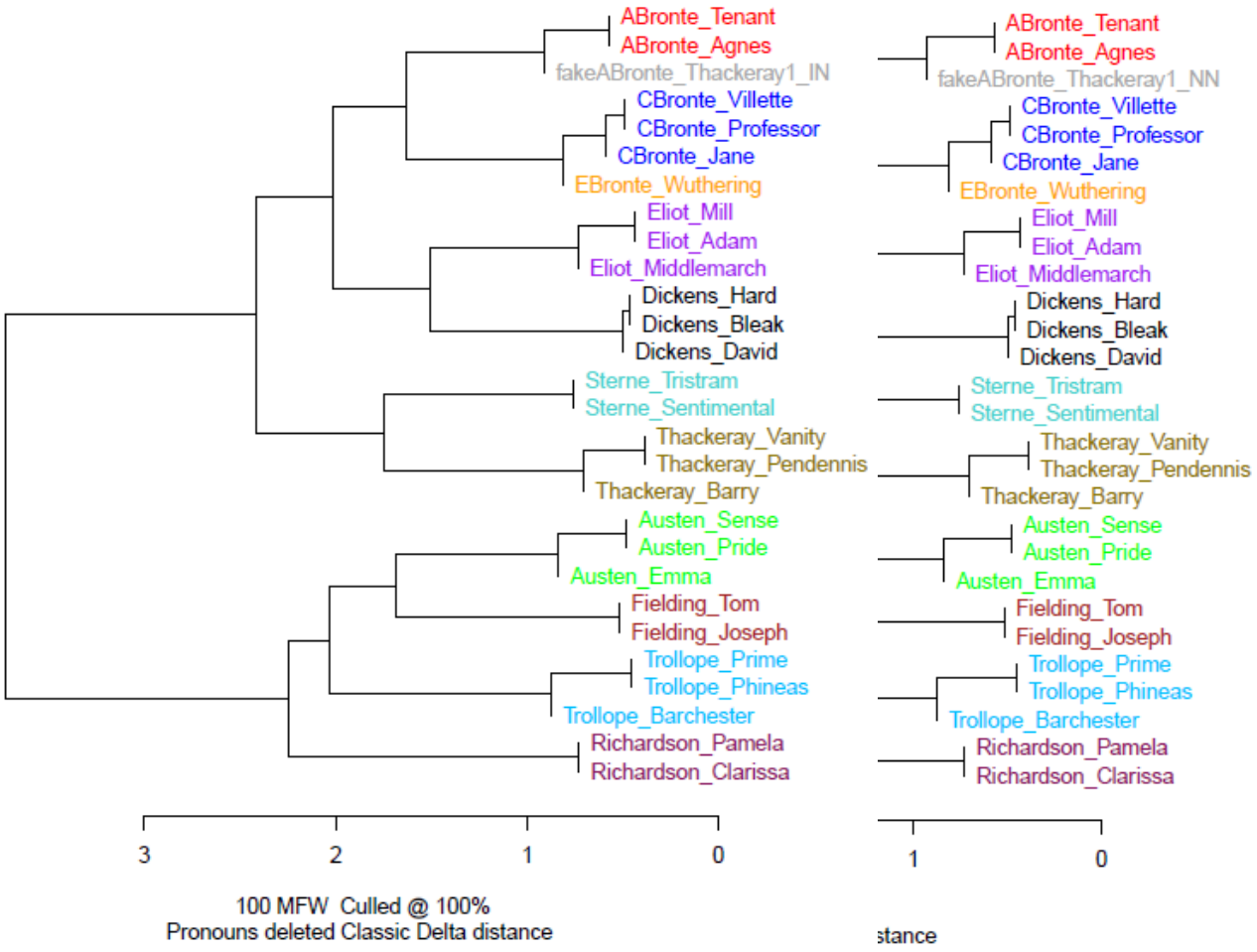

Figure 4 Dendrogram of classic British novels together with the computer-generated hybrid text based on A. Brontë's PoS-tag distribution and the frequencies of words in IN or NN categories, respectively, and the rest being Thackeray's vocabulary. The hybrid texts were clustered together with Anne Brontë's other novels (note that there are only 100 MFW taken into account).

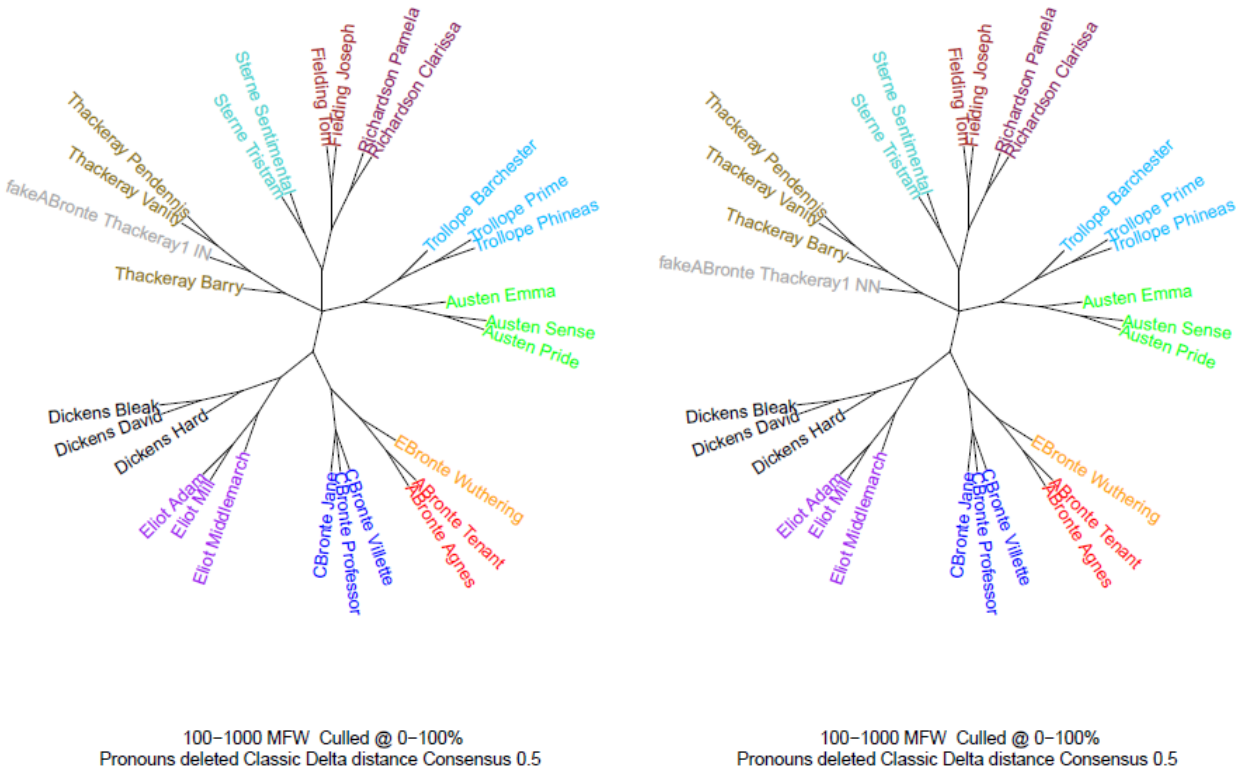

Figure 5 Consensus trees of classic British novels together with the computer-generated hybrid text based on A. Brontë's PoS-tag distribution and the frequencies of words in IN or NN categories, respectively, and the rest being Thackeray's vocabulary. The hybrid text stays on the Thackeray's branch for 100-1000 MFW, but the substitution of nouns is powerful enough to make it the least Thackeray-ish one.

2.3. Results: Translatorial traces or the influence of the original language

In this chapter, the material consists of a series of novels written by a single author and translated by a single translator only, hence no conclusions can be drawn as regards distinctness of literary translations from the rest of literature in the target language (provided that there is any in general); instead, I attempt to examine whether the translations under scrutiny are in any way untypical by comparison with the native, target language literature. One would expect that translators are very much aware of their language use and can consciously and meticulously screen the foreign language interference that could manifest itself as calques; they are trained to avoid them. Furthermore, translated novels are also edited, proofread, etc., to ensure correctness according to the target language norms. Lack of calques and language errors, however, does not exclude language transfer in the form of a systematic bias towards this or other correct grammatical structure. Thus, my focus is on what is encoded in the basic grammatical information, namely frequencies of parts of speech, so that one could trace systemic differences from the norm. Whether they be caused by the influence of the language of the original, or the style of the original, or the style of the translator, is not easy to judge if at all possible.

To proceed with the study I began with constructing a common part-of-speech tag-set for English and Polish, as described in **Appendix B.1**, in order to compare tag distributions in the originals and translations, in case it could provide us with a clue as to where any language transfer could be expected. I believe more sophisticated tag-sets are needed for that purpose than the 12-PoS universal tag-set introduced by Petrov et al. (2012), which, incidentally, still lacks a mapping from the Polish language tree banks. It might also be the case that searching for language transfer in translated texts *might* involve a different tag-set than just the one used for the purpose of parsing if one already expects some specific syntactic structures to be systematically rendered by different parts of speech.

Let us first proceed with look at how merely the tag distributions for PL and EN differ, see **Figure 6**. The full explanation of the tag meanings can be found in Santorini (1990), and some necessary modifications are described in **Appendix B.1.** In short, the most common tags are: nouns (NN) and proper nouns (NNP), various forms of verbs and participles (starting with V), adjectives (JJ) and adverbs (RB) in different degrees, personal pronouns (PRP), coordinating conjunctions (CC) and subordinating conjunctions together with prepositions (IN). The differences in the tag distributions are not of much interest to us, however; they serve us only as a reference when we go on to study translations.

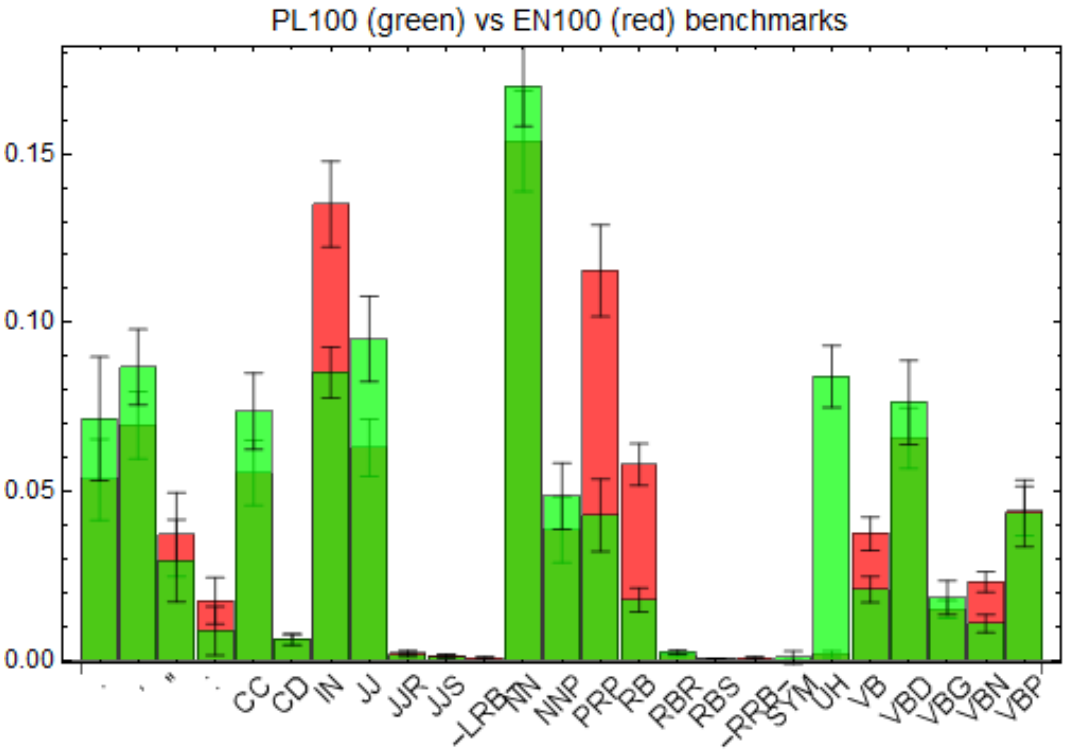

Figure 6 **The differences between PoS-(and punctuation)-tag distributions often come from different tagging schema for the two languages, e.g., putting together subordinating conjunctions and prepositions (IN) in English.** Specific comments can be found in **Appendix B.1**. The error bars represent bookwise standard deviations of the number of tokens in a given PoS class.

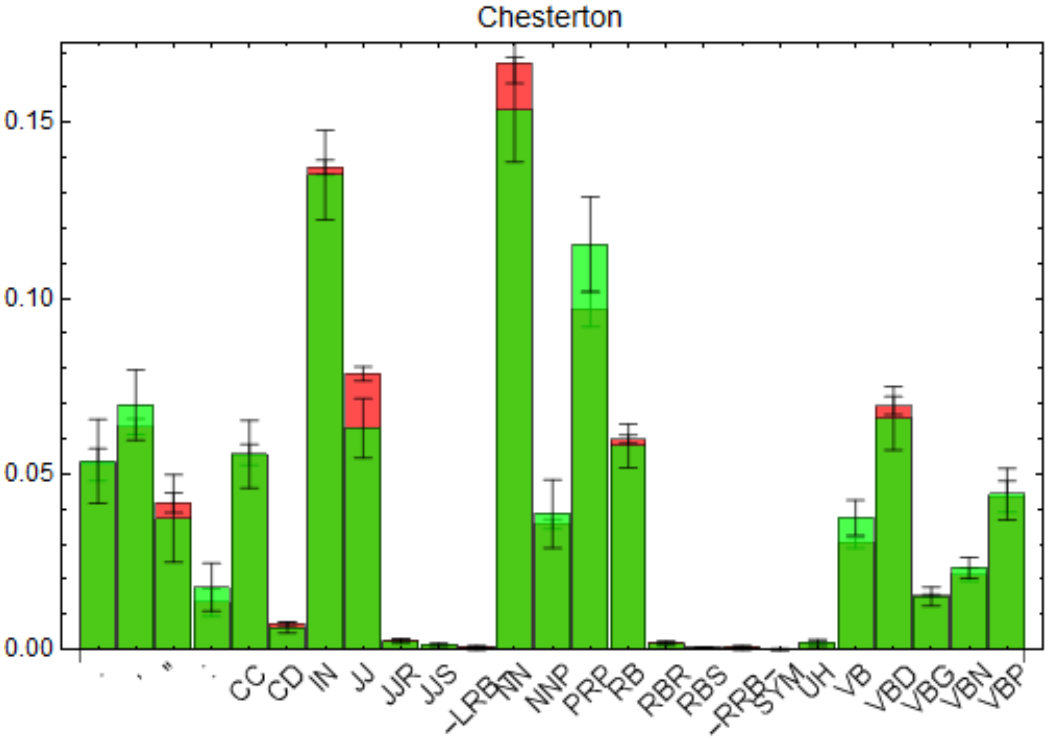

Figure 7 **An example of how Chesterton's PoS-tag distribution (red) differs from the EN100 corpus mean (green).**

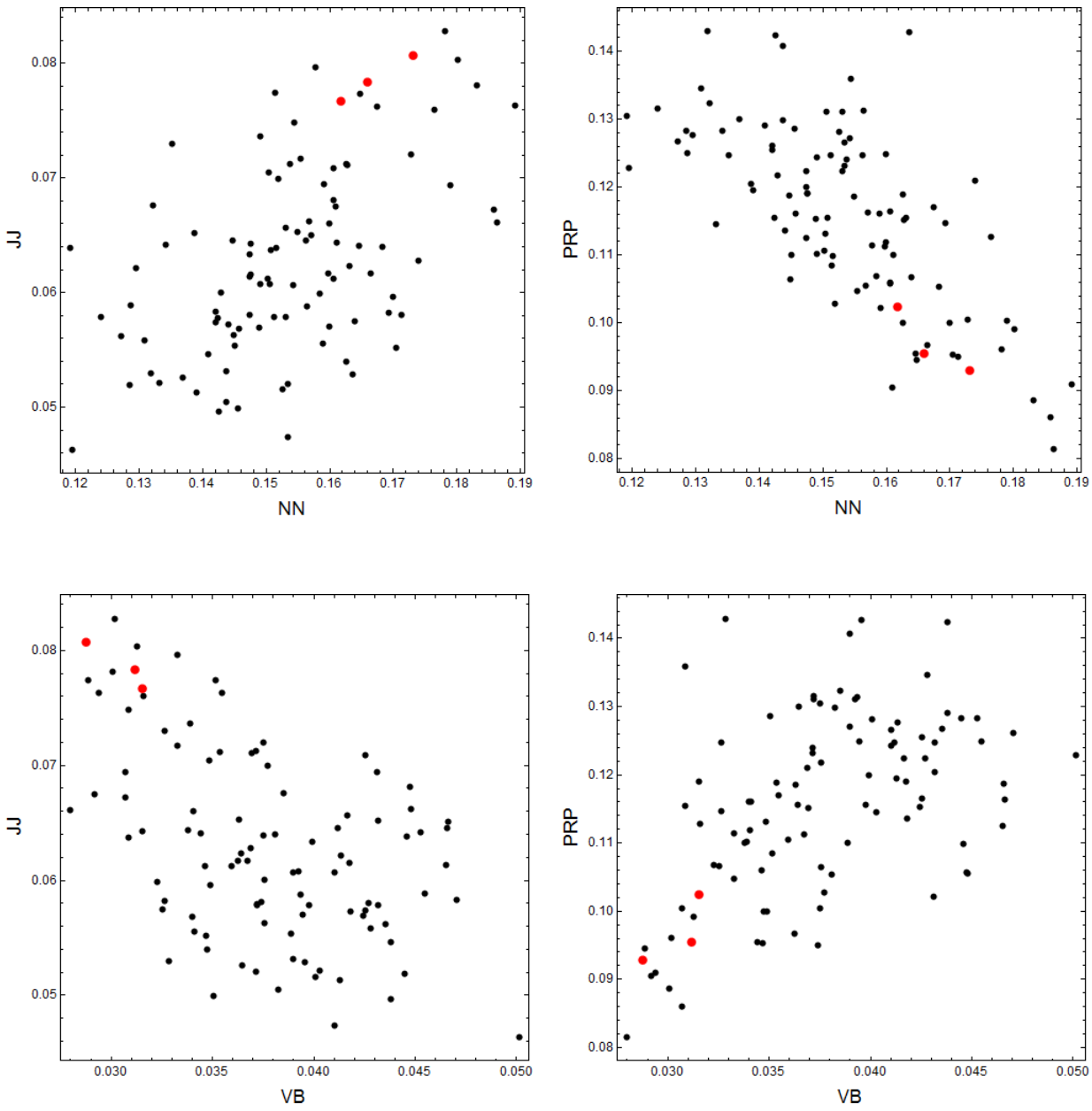

Figure 8 Example correlations between PoS tags. Each point represents a book in EN100 corpus. The three red points correspond to works by Chesterton (cf. Figure 7). The absolute values of Pearson correlations between pairs (JJ, NN), (PRP, NN), (PRP, VB) are slightly more than 0.5, (VB, NN) almost 0.6, and (PRP, NN) is 0.7. Note that this is a very coarse-grained behaviour on the scale of whole books.

The more interesting information is how the distributions of particular authors vary with respect to the mean distribution of the whole corpus. This is illustrated in **Figure 7** with collective distribution for three works of Chesterton's, where, e.g., noun, adjective, pronoun, and verb frequencies are the ones most diverging from the mean. As noted in Argamon (2008), one has to be careful with correlations between such deviations; indeed, it appears that there are some correlations and anticorrelations between these parts of speech, as can be seen in **Figure 8**, which means that to some extent increased number of nouns already accounts for increased number of adjectives and decreased number of verbs

and pronouns. This in turn, could be expected to a certain extent from superficial knowledge of grammar (such as: “adjectives modify nouns”). A more detailed connection of these collective deviations to style, however, is missing.

The particular English-to-Polish translation example can perhaps reveal something more intriguing. Below I analyse the corpus of *Discworld* fantasy novels by Terry Pratchett and their translations, all by the same Polish translator, Piotr Cholewa.

As shown in **Figure 9**, the number of full stops, quotation marks, and expletives is more than one standard deviation higher than in the original; the number of the subordinating and coordinating conjunctions is, as expected from the above, much lower, and possibly present, past, and gerund verb forms are slightly more frequent than in the benchmark corpus. Very generally speaking, this might indicate a larger proportion of short dynamic sentences.

Now, what happens in translation is reproducing the pattern of deviations from benchmark: by visual inspection all directions (i.e., higher or lower than average) are retained but for VBG and quotation marks. In the former, it might be mistagging the gerunds in English and an inappropriate design of the common tag-set (where I put in VBG Polish gerunds as well as some of adjectival and adverbial participles, whose usage is distributed somewhat differently than English gerunds and present participles). In the latter, the culprit is different typographical conventions in Polish, where character utterances in dialogues begin with a dash (as confirmed by larger count in the “**:**” tag class), and unfortunately, some tagging errors.

For now, the results show that the translator probably followed the structure of the original text, but probably did not break out from the distribution of tags expected in Polish texts in general, as comparison with the Polish corpus shows (perhaps with the exception of suspicious overuse of punctuation – checking sentence lengths would involve some additional examination).

An additional check over the joint behaviour of pairs of tags, **Figure 10**, shows some unexpected tendencies.

Although these considerations do not yet reveal the translatorial component measured by authorship attribution methods (assuming there is any), they already allow to detect some fishy behavior.

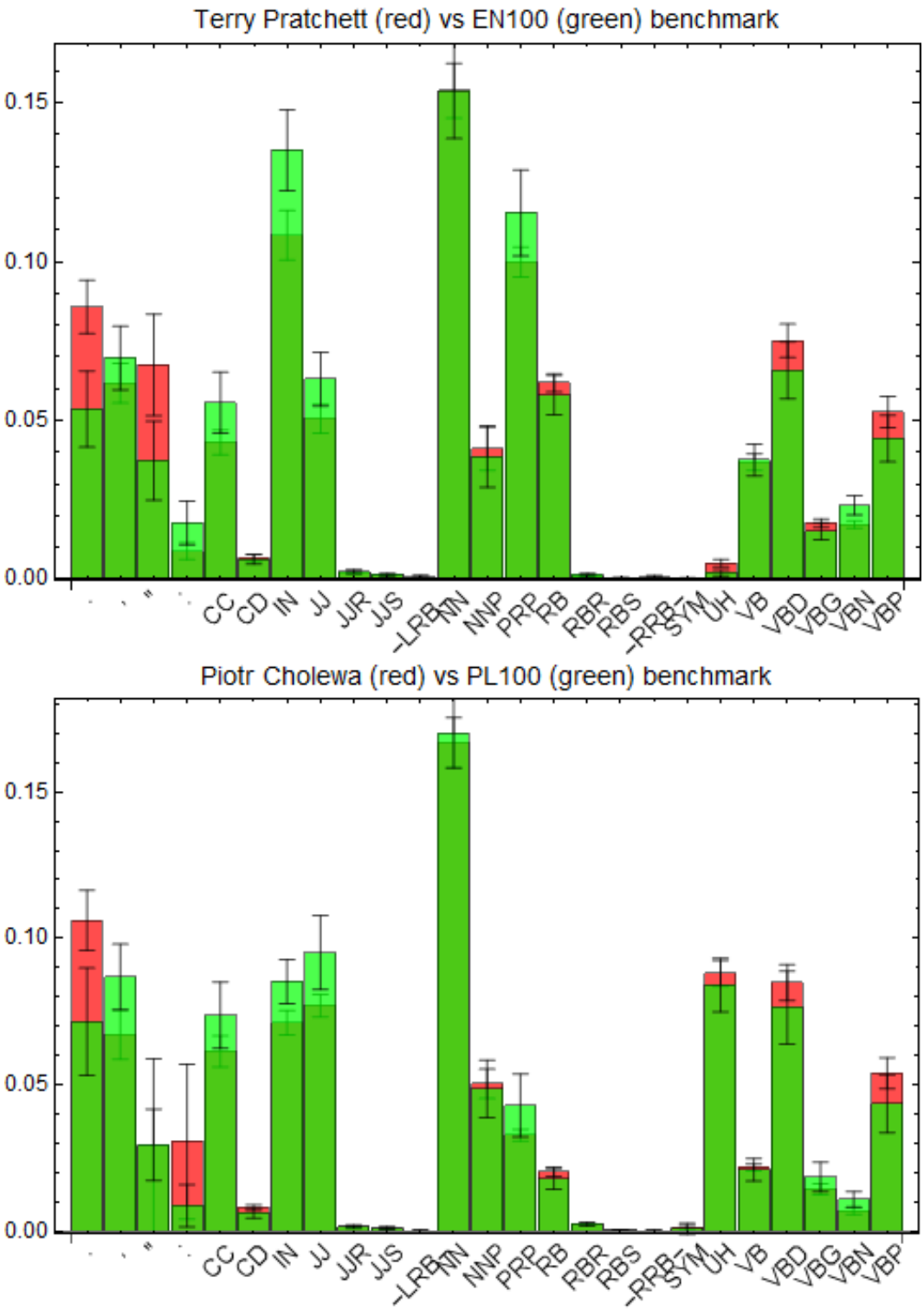

Figure 9 Tag distributions in the original and translated (EN -> PL) *Discworld* novels compared to respective benchmark distributions.

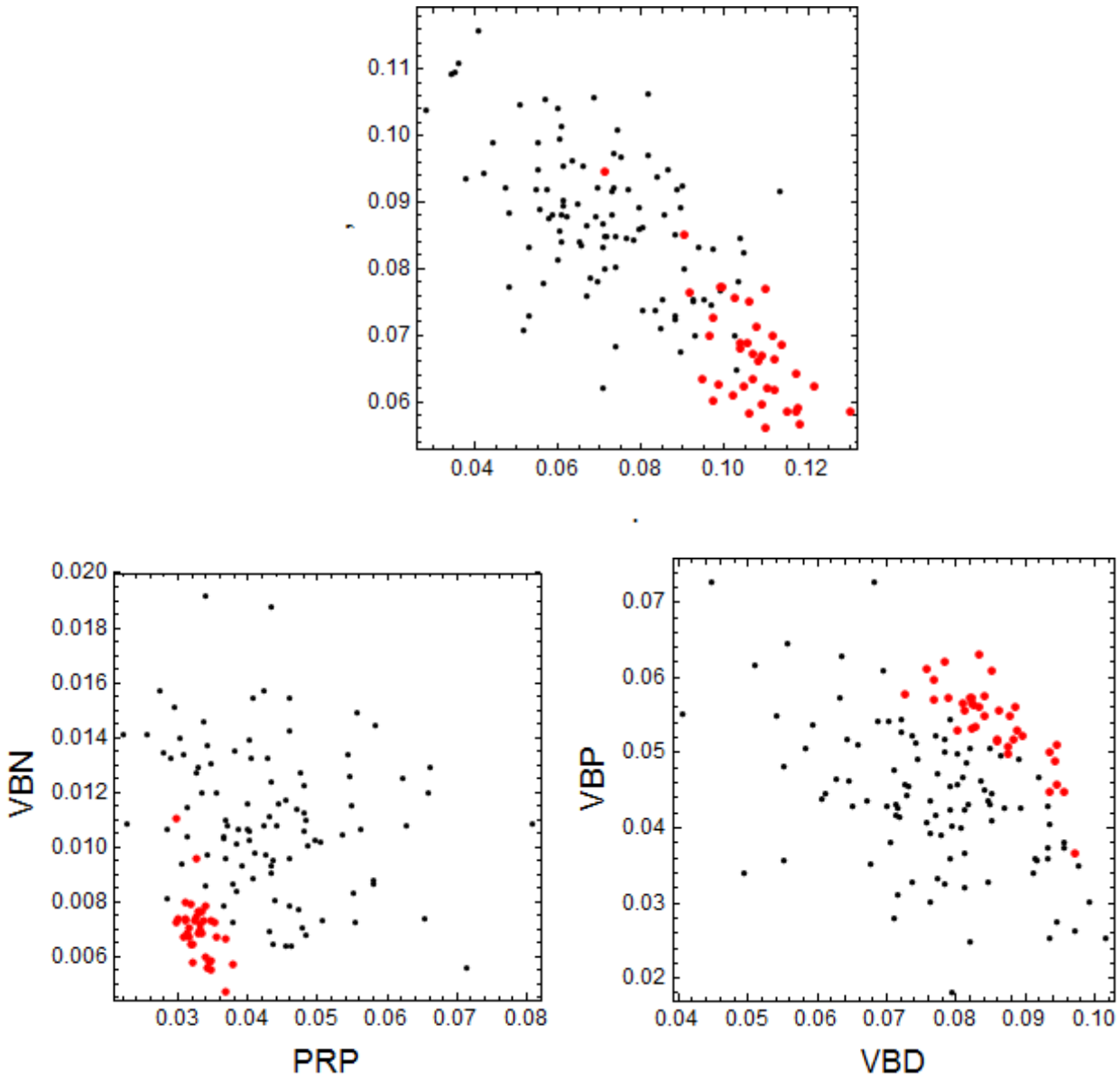

Figure 10 **Unnaturally correlated tags in Polish translation of *Discworld* (red) as compared to PL100 benchmark (black). The overall punctuation (upper panel: frequent sentence-final punctuation and infrequent commas) seems to be at the verge of what is normal in Polish. Similarly, rather scarce past participles and pronouns (bottom left) and simultaneous fairly abundant past and present tense forms, jointly indicate a style stretching at the limits of the normal usage.**

There can be no community between you and me; we are enemies. Begone, or let us try our strength in a fight, in which one must fall.

Mary Shelley
Frankenstein or The Modern Prometheus, 1818

3. Using methods of community detection to attribute authorship

In this chapter I present an alternative to the standard classifiers and clustering algorithms utilized at the very end of the procedure chain attributing authorship – the so called community detection methods for complex networks [see Fortunato (2010) for a broad review]. They are a family of unsupervised methods, which means that they do not have a training sample and so generally do not fall into the trap of overtraining/overfitting. Only some kind of overtraining may take place, if these methods are parameterised; then, different optimal parameters may exist for different corpora. In this case, the resolution of the modularity null-model is such a parameter (Fortunato and Barthelemy 2007). This issue can be remedied to some extent, as there are some resolution-free methods (Traag et al. 2011), still to be tested on the linguistic data.

Eder and Rybicki (2012) brush aside the unsupervised machine learning techniques, saying that "they require human interpretation of the degree of similarity between analysed samples" and so "are subjected to the attributor's arbitrary decisions." I hope that the examples I provide below take a stand against that view. In fact, the problem discussed in the paper quoted is about the large and highly unpredictable dependence of the results on training sample for the supervised methods – a matter non-existent in the unsupervised method I present.

In fact, as mentioned in Craig (2004), they do not really need bootstrapping. The methods of community detection compare the data with an in-built null model of what the whole table of connections should look like instead of constructing a model of each author from the training sample and then comparing all the test texts with these authorial models. Thus, needing just one iteration, these methods might be computationally faster.

The algorithms presented here exist also in more sophisticated versions, where so called "overlapping communities" are allowed, i.e., some items can be classified into several groups. This is usually not welcome in authorship attribution, where we aim at finding the one and only author, but is similar to giving several best scores in cross-validation procedure. The standard hierarchical clustering algorithms can be additionally enhanced with a bootstrap consensus technique providing a more stable and reliable results (Rybicki 2011); this might be possible for the network-based (as opposed to tree-based clustering) community detection methods as well.

3.1. Methods: graphs and clustering

The clustering method presented here has been devised for clustering of *graphs* (in the sense of the graph theory), i.e., abstract entities which comprise so called *vertices* and *edges* (also called *nodes* and *links*), as exemplified in **Figure 11**. One of the virtues of such graphs is that they can be rewritten numerically in the form of matrices that encode the connections between the nodes. One of such matrices is the *adjacency matrix*, which for the graph in **Figure 11** yields:

$$\begin{bmatrix} 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0.3 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 \end{bmatrix},$$

where the first row says that node A is not connected to itself or to node D (zeros at positions 1 and 4) but it is connected to nodes B, C, and E; the second row shows that node B is connected with weight 0.3 to node C; and so on.

Now, it can easily be seen that the table of distances Δ in point IV of **Section 1.2.** is very similar to such adjacency matrix. The difference is that the adjacency matrix rather shows similarity between nodes, while Δ shows distances, which roughly is a reciprocal: if two objects are not similar to each other, they are distant. The point is that from Δ one can construct a graph, and so, seemingly, one could utilize algorithms for clustering analysis of graphs.

One of the problems, however, is that the distance matrices produced by Burrows's Delta are not sparse, quite the contrary, all but the diagonal elements are non-zero. In the graph representation this is a *weighted complete* graph, i.e., briefly, everything is connected to everything else. In such situation, the distance tables might need some pre-processing. Consider the extreme case, in which Delta distance from author A to him- or herself is 0; the inverse, the similarity, would be infinity; but the precise functional form of the relation between distance and similarity matrix, can be chosen fairly arbitrarily, which then can affect the final results. The chosen function is described in **Appendix B.5**.

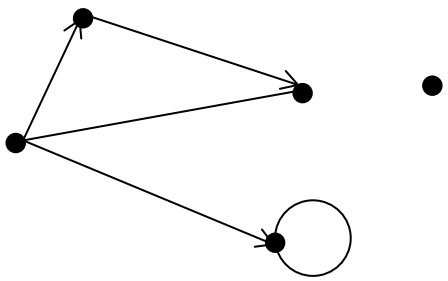

Figure 11 Exemplary graph with 5 nodes (A, B, C, D, E), 1 undirected and unweighted link {A,C}, 1 undirected multiedge loop ({E,E},2), 2 directed unweighted links {A,B} and {A,E}, and 1 directed weighted link ((B,C),0.3).

3.2. Results: community detection algorithms versus the Delta

Once we know the representation that is fed into the algorithms, one cautionary note is in place: Burrows's Delta produces graphs which are complete. In fact, it appears that such structure of the data makes one of the best methods of community detection, Infomap (Edler and Rosvall 2013), fail. The method I use further on is the Louvain method of modularity maximisation as presented by Blondel et al. (2008). I compare these methods only on English language corpora, because the use of other languages, just as of other corpora, only changes the Delta distance table, which is an independent step before any clustering algorithm comes into play, while in this chapter the change in the method only involves taking a different clustering algorithm.

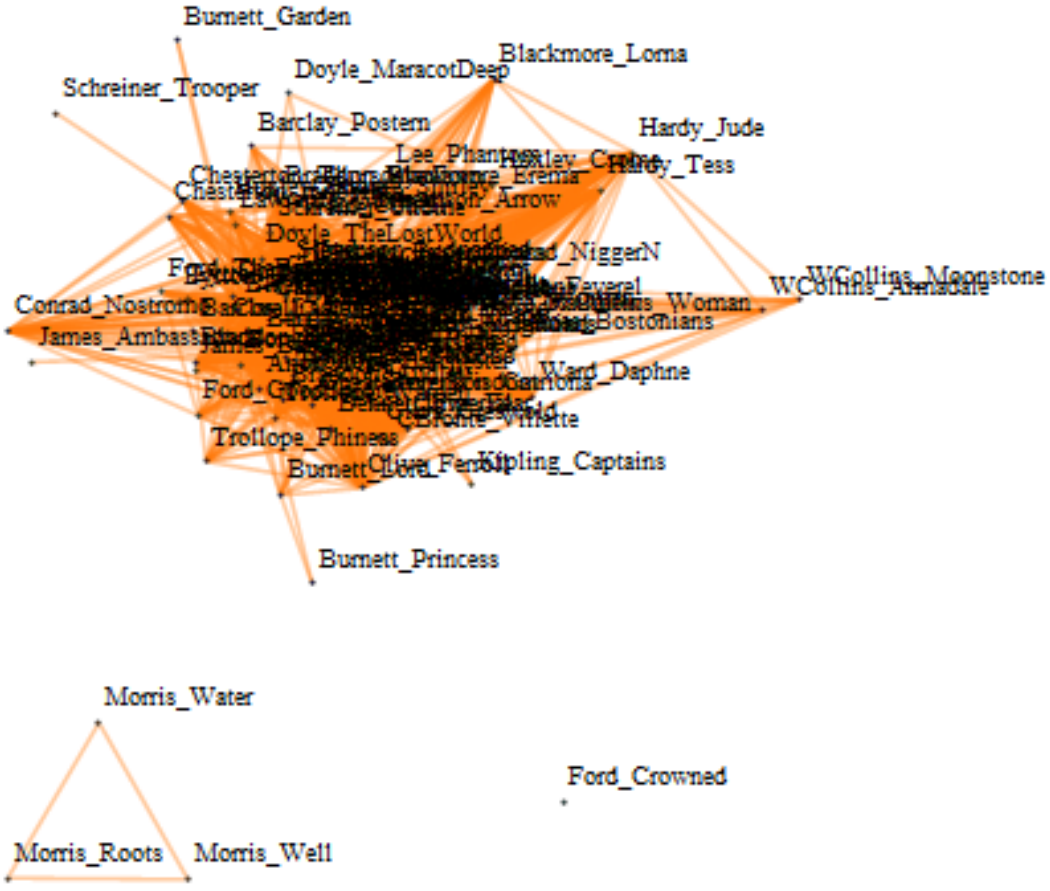

Figure 12 Very dense (almost complete) graph produced from the table of Burrows's Delta distances for the benchmark corpus of 100 EN novels. The visualization contains virtually no information.

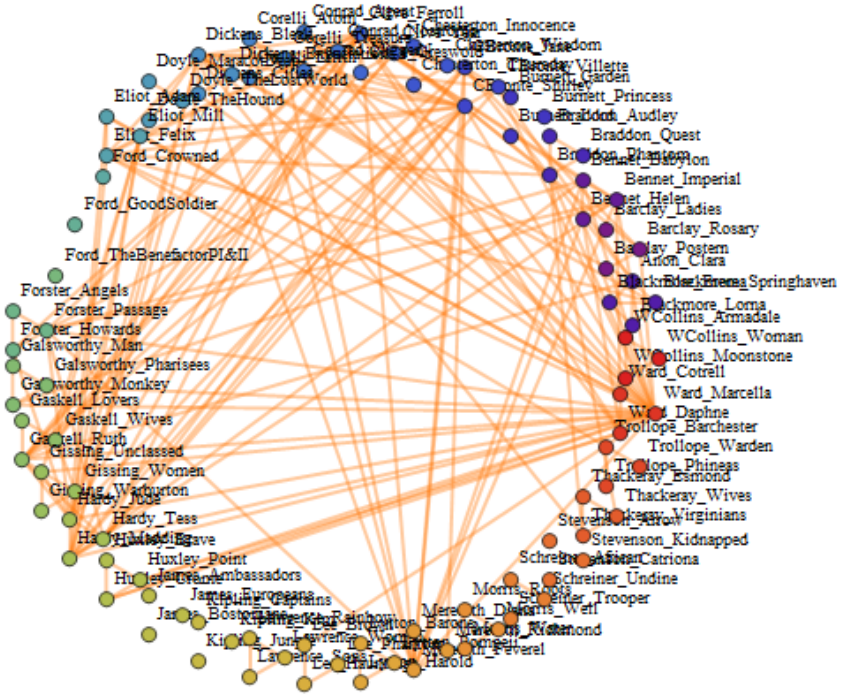

Figure 13 Visualisation of the benchmark corpus of 100 EN novels utilising the information obtained from clustering (Blondel et al. 2008): each colour marks books considered by the algorithm to be authored by the same person; such groups of 1-4 books were aligned close to each other. Only the strongest links in the graph are shown for visual clarity (the clustering used all the links).

As can be seen in **Figure 13**, the clustering of novels in EN100 benchmark according to the Louvain method of community detection works well. The usual accuracy of state-of-the-art methods is around 96-98%, as shown in **Figure 14**.

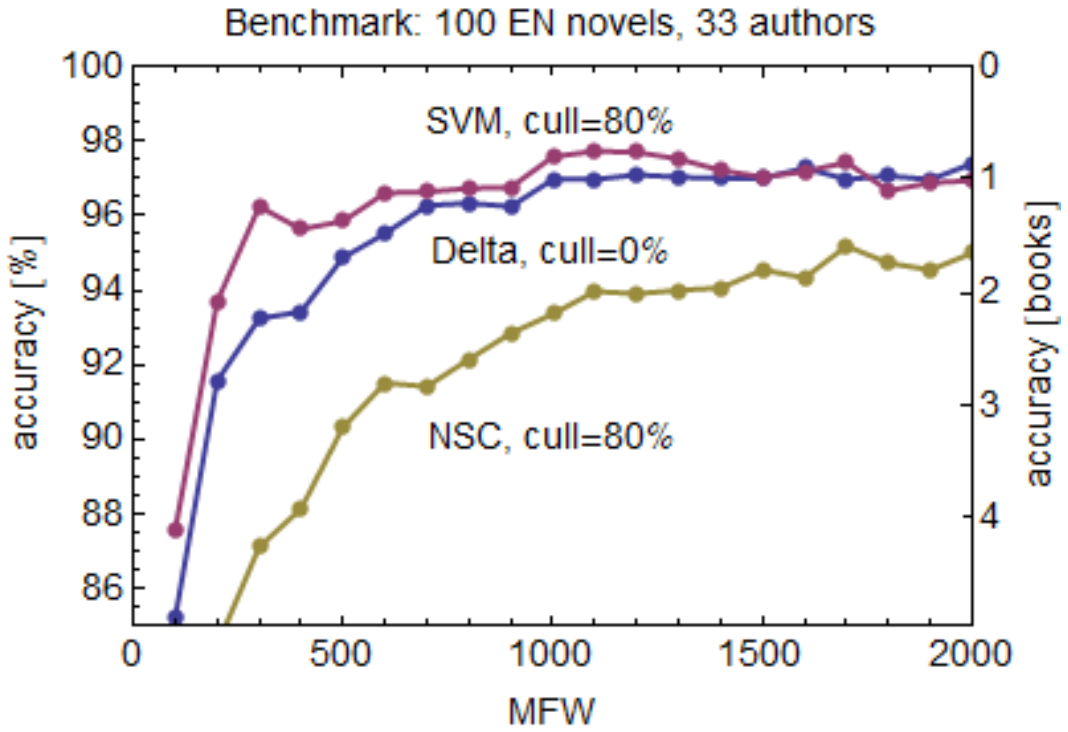

Figure 14 Results of authorship attribution on the EN100 benchmark with the use of Burrows's Delta, Nearest Shrunken Centroids, and Support Vector Machines [as implemented by Eder et al. (2013), after 100-fold cross-validation.

Actually, there does not exist an easy comparison between the results obtained by the supervised methods and the unsupervised community detection methods. The supervised ones, provide a Yes/No answer to the question "Does this book is similar to the other books written by A?". In the community detection methods, we do not provide any training information saying "These books were written by A." The only thing they do is finding groups of similar books, so the results they provide is: "Books no. 1, 7, 8 are in the same group; no. 2,3,4 are in the same group;" and so on; there is no a priori authorial label of the group. This might result for instance in two books of one author and two books of another author forming one group – but then which books should we consider misclassified? The first two, the second two, or all four?

The way to approach it is to use *Normalised Mutual Information* (NMI) [first used in the context of comparing clusterings by Danon et al. (2005), which can compare the true authorial groups (which are known for the benchmark) with the groups obtained from clustering methods on information-theoretical grounds; it takes values from 0 (the true and found groups are totally unrelated) to 1 (they are identical), but its relation to the percentage of misclassified entities is non-linear and thus the value of NMI should not be interpreted as any such percentage.

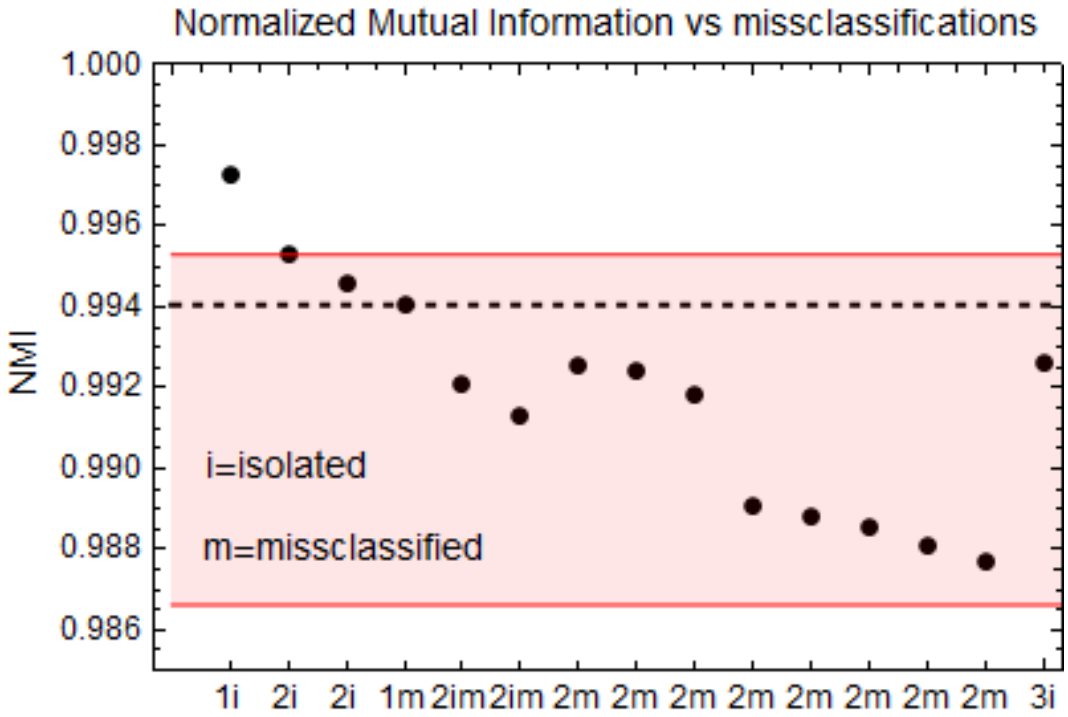

Figure 15 Example of what is the value of Normalised Mutual Information for different types of misclassifications on EN100 benchmark corpus. The dashed line is the value obtained by the state-of-the-art authorship attribution methods (1 misclassification in the 33 authorial groups). The pink-shaded region is the extent of NMI values for different possible ways of misclassifying two books. In Figure 13, the community detection method separated the three books by Ford (upper left), which corresponds to isolating two books, labelled 2i (the 2nd point from the left).

Although **Figure 15**, showing results for the EN100 corpus indicates that the community detection methods might have a slight edge over Delta, **Figure 16** for a larger EN500 benchmark dispels my illusions about unsupervised methods –based on NMI (i.e., even knowing the true authorial groups) it is even hard to say how many authors there are, as there is no clear NMI maximum. In this figure, the different numbers of clusters (authorial groups) were obtained by changing the resolution parameter in the null model of modularity.

It has to be stressed, nevertheless, that in this comparison the supervised methods implemented in Stylo actually attempted classification of 100 books knowing the true authorship of the remaining 399 novels. So if they attempted to look at another 100 out of 499 novels, and yet another 100, scoring the 87% accuracy each time, it would mean that there are in fact around 5*13 = 65 novels that pose problems. The Louvain method on the other hand had no prior knowledge: all the clusters are emergent.

Further development of the research summarised in this chapter could be the inclusion of supervision into the existing community detection methods, which as it seems has not been done.

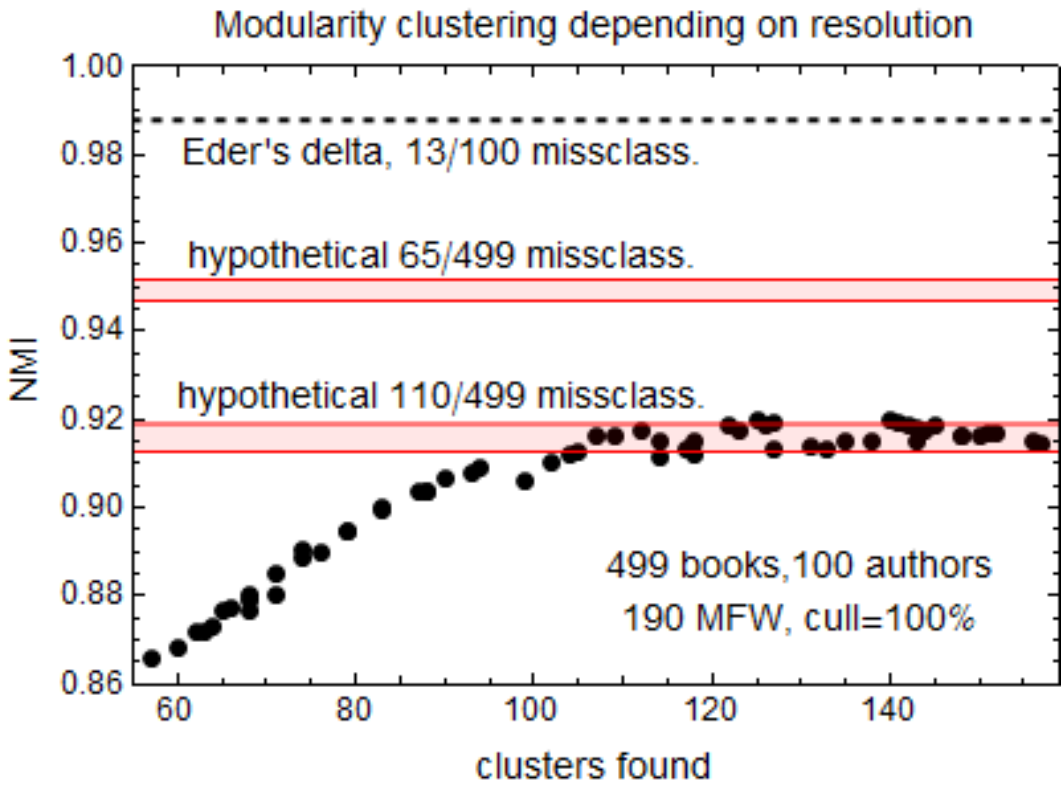

Figure 16 The black points are the result of clustering of EN500 corpus obtained by the Louvain method for increasing resolution. The plateau corresponds to roughly 110 books having incorrect group membership. The result for a supervised classifier (399 books for training; 100 for test) is plotted with the dashed line. 190 was the maximum number of most frequent words that all the 499 novels have in common.

The precise results of clustering the 499 novels is:

- correct grouping of all the books by: Anderson, Charlotte Brontë, Emily Brontë, Brown, Cather, Chesterton, Christie, Clancy, Coben, Collins, Disraeli, Eliot, Fitzgerald, Forster, Freeman, Gaskell, Glasgow, Grisham, Hall, Hardy, Kipling, Koontz, Lawrence, Lewis, Ludlum, Lytton, MacDonald, Mansfield, Mason, McNeile, Meredith, Morrison, Post, Powys, Dorothy Richardson, Rowling, Scott, Thackeray, Trollope, Waugh
- Doyle, Galsworthy, Greene, Hawthorne, Huxley, Lessing have each been split into 2 groups
- Dickens and James have each been split into 3 groups
- Compton has been split into 2 groups, and absorbed one of Conrad-Ford collaborations
- Conrad has been split into 2 groups, and absorbed one of Conrad-Ford collaborations
- Ford has been split into 5 groups, and absorbed one of Conrad-Ford collaborations
- Melville has been split into 2 groups and absorbed one of Twain's books
- All Morris's books but one (*Signs of Change*) have been grouped together
- All Orwell's books but one (*Animal Farm*) have been grouped together
- All Edgeworth's books but one (*Castle Rackrent*) have been grouped together
- All Maugham's books but one (*Liza of Lambeth*) have been grouped together
- Fielding's *Shamela* has been grouped with Richardson
- Mitchell's *Gone with the Wind* has been grouped with Montgomery
- Golding has been split into 2 groups, and absorbed Passos
- Anne Brontë has been grouped with Jane Austen
- Obrien has been grouped either with Stephens:
    1. Obrien_Policeman, Stephens_Crock, Stephens_Mary

or with Joyce, who has been split into 2 groups:

1. Joyce_Finnegans, Joyce_Ulysses, Obrien_Swim
2. Joyce_Dubliners, Joyce_Portrait

- Tolkien and Nabokov have been grouped correctly but for one novel each (*Silmarillion* and *The Real Life of Sebastian Knight*, respectively), which have been grouped together with all the books by Wyndham Lewis, Wells, Wharton, Wilde, Woolf
- Stevenson's group has absorbed Twain's *The Adventures of Tom Sawyer*, *The Prince and the Pauper*, and *A Connecticut Yankee in King Arthur's Court*
- Three of Faulkner's novels have been grouped together; the other two are grouped with a bag of Green's, Hemingway's, Maugham's, Salinger's, Stein's, Steinbeck's, and Twain's

The rest of the books had formed rather incomprehensible groups. Until I started compiling the list of authors and titles – it appeared to me that some of the writers that it is really hard to find (on Wikipedia) are the woman writers related to the so called Chawton House, and indeed most of them group together into a few clusters. Parenthetically, since I did not have the text files of the 499 books (only the filenames and the table of Delta distances) it made the work a little bit easier, as I could directly google the Chawton House website instead of struggling through the thicket of irrelevant results from the whole Internet.

All of the groups resulting from the clustering are given in **Appendix A.3**. Such results might indicate that the method might have problems with the resolution, i.e., with the varying group sizes, which leads to splitting authorial groups. The other inconsistencies might already be indicative of some similarities of authorial style (vide *Shamela,* or joining Austen with A. Brontë).

In the above I have chosen the resolution of modularity which yielded the number of groups closest to the hypothetical 100 (which means that there was some auxiliary information needed).

I had committed deeds of mischief beyond description horrible, and more, much more
(I persuaded myself) was yet behind. […] I had been the author of unalterable evils, and I lived in daily fear lest the monster whom I had created should perpetrate some new wickedness.

Mary Shelley
Frankenstein or The Modern Prometheus, 1818

Discussion

The results presented in **Chapter 2** prove that the grammatical and lexical layers of authorial style are separable, at least to some extent, and that the Burrows's Delta measure of distance between texts mixes these two layers in different proportions depending on the range of words chosen as the basis of the comparison. It has been shown that preserving merely the distribution of parts of speech (i.e., simply the relative numbers of words belonging to the categories) and the vocabulary comprising one of PoS classes is enough to fool the authorship attribution algorithm if it uses only the 100 most frequent words. How the preservation of PoS distribution can be carried out in real life (i.e., preserving the sense of the text) shrouded in mystery. What lies beneath the shroud is all the different kinds of sentence, clause, and phrasal structures that can aggregate into the distribution characterising one or the other author. Further studies on that topic should involve parsing texts to obtain syntactic dependencies or at least some advanced analysis of PoS n-grams. The lexical layer also needs to be further split at least into topic-specific word groups, so that the topics of individual texts could be abstracted from the authorial style. The results shown in this thesis also need to be repeated for other pairs of authors, and then, even further for groups of authors with the age, genre, topic and other variables controlled for.

The part of the chapter that was concerned with translation used a different method based solely on the distributions and co-occurrences of PoS tags, which is due to the fact that it was rather the grammatical rather than lexical layer that I assumed could show more language transfer. Since the translator had to retain all the vocabulary specific to the fantasy novels genre, to the *Discworld* series, and to the particular book, the grammatical structures were assumed to be the significant distinguishing factor that could be prone to some language transfer. Such influence was observed for the idiomatic language in the novels (Hantz 2013), which however seems too little to affect the average PoS distributions. The problem here was that the benchmark corpus used was of a general type, while perhaps a corpus of native Polish fantasy literature could be used to control for the subgenre. The crude methods, nevertheless, do indicate some suspicious correlations between several parts of speech, which could be traced semi-manually. The drawback of the correlation plots is also that they are very coarse-grained, because each point in them represents one book, which means that some internal averaging has already been performed; perhaps it could be alleviated by subsampling or chopping the novels into smaller parts. The authorship attributions methods were not used for this part of the study, so that although I may have gained some insight into what does distinguish translations from the literature in the target language, I did not proceed into checking how this affects the authorship attribution of translations.

Next, in **Chapter 3** I shortly introduced how the methods of community detection in complex networks can be applied instead of the classical clustering and machine

learning techniques. Since the performance of the methods changes together with the corpus under investigation, there still awaits an extensive systematic study to be done before a fair comparative judgement can be passed. Choosing benchmark corpora for testing performance of clustering methods, however, can be very misleading, as the input information for clustering is only the distance table (or adjacency matrix in the graph-theoretical interpretation) and it is rather the benchmark distances (graphs) that should be used for comparison between those. In benchmark corpora one should not count on 100% correct authorship attribution due to the genre, topic, gender, pastiche, collaboration, and other effects which can obscure the authorial fingerprint, especially when the most frequent words (or n-grams) are taken indiscriminately. The choice of the features for the analysis is outside the jurisdiction of the clustering methods, although some other machine learning methods might perform feature selection by themselves. It should also be reiterated that the community detection algorithm I used in this thesis return a partition of texts without supervision, i.e., it proposes that some texts seem to form groups without actually comparing them to any particular authorial prototypes.

Finally, let me briefly conclude, based on the above results, why stylometry and authorship attribution work and what they measure; the answer to the first question, however unsatisfactory, probably should read as follows: they work well because they measure many aspects of texts simultaneously without discriminating between them (which is, by the way, also an answer to the question why authorship attribution occasionally fails). This immediately leads to the second answer, which is almost as uninformative – the greatest failure of this thesis, – the Burrowsian way of doing stylometry and authorship attribution blends the grammatical, lexical, and too many other textual layers by indirect measurement of sentence structures and narration (the distributions of punctuation, conjunctions, pronouns, etc.), of phrasal structures (the relative statistics of most of parts of speech), the topic, content and the semantic domains (all the content words), the genre and age (encoded already mostly in all the mentioned above), and so on. **Chapter 2** seems to provide methods to arrive at more precise statements concerning the relative contributions (depending on the range of features) of each of these layers, and in the long run could help to decouple them.

Bibliography

Philip W Anderson et al. More is different. *Science*, 177 (4047): 393–396, 1972.

Shlomo Argamon. Interpreting Burrows's Delta: Geometric and Probabilistic Foundations. *Literary and Linguistic Computing*, 23 (2): 131–147, 2008. doi: 10.1093/llc/fqn003. URL http://llc.oxfordjournals.org/content/23/2/131.abstract.

Mona Baker. Towards a methodology for investigating the style of a literary translator. *Target*, 12 (2): 241–266, 2000.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008 (10): P10008, 2008. URL http://stacks.iop.org/1742-5468/2008/i=10/a=P10008.

John Burrows. 'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing*, 17 (3): 267–287, 2002a. doi: 10.1093/llc/17.3.267. URL http://llc.oxfordjournals.org/content/17/3/267.abstract.

John Burrows. The Englishing of Juvenal: computational stylistics and translated texts. *Style*, 36 (4): 677–699, 2002b.

John Burrows. Textual Analysis. In Susan Schreibman, Ray Siemens, and John Unsworth, editors, *A Companion to Digital Humanities*. Blackwell, Oxford, 2004. URL http://-www.digitalhumanities.org/companion/.

John Burrows. All the Way Through: Testing for Authorship in Different Frequency Strata. *Literary and Linguistic Computing*, 22 (1): 27–47, 2007. doi: 10.1093/llc/fqi067. URL http://-llc.oxfordjournals.org/content/22/1/27.abstract.

Chawton House Library. URL http://library.chawton.org.

Chawton House Library. Introductory note. In Penelope Aubin, *The Life of Charlotta Du Pont, an English Lady*. URL http://www.chawtonhouse.org/wp-content/uploads/2012/06/The-Life-of-Charlotta-Du-Pont-an-English-Lady.pdf. [accessed: 10 September 2014]

Hugh Craig. Stylistic Analysis and Authorship Studies. In Susan Schreibman, Ray Siemens, and John Unsworth, editors, *A Companion to Digital Humanities*. Blackwell, Oxford, 2004. URL http://www.digitalhumanities.org/companion/.

Leon Danon, Albert Diaz-Guilera, Jordi Duch, and Alex Arenas. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005 (09): P09008, 2005.

M. Eder, M. Kestemont, and J. Rybicki. Stylometry with R: a suite of tools. In *Digital Humanities 2013: Conference Abstracts*, pages 487–489, Lincoln, 2013. University of Nebraska-Lincoln. URL https://sites.google.com/site/computationalstylistics/preprints/Eder-Kestemont-Rybicki_Stylometry_with_R.pdf.

Maciej Eder. Personal communication, July 28, 2014.

Maciej Eder. Does size matter? Authorship attribution, small samples, big problem. *Literary and Linguistic Computing*, 2013. doi: 10.1093/llc/fqt066. URL http://-llc.oxfordjournals.org/content/early/2013/11/14/llc.fqt066.abstract.

Maciej Eder and Jan Rybicki. Do birds of a feather really flock together, or how to choose training samples for authorship attribution. *Literary and Linguistic Computing*, 2012. doi: 10.1093/llc/fqs036. URL http://llc.oxfordjournals.org/content/early/2012/08/10/-llc.fqs036.abstract.

Maciej Eder. Style-markers in authorship attribution: a cross-language study of the authorial fingerprint. *Studies in Polish Linguistics*, (6), 2011.

D. Edler and M. Rosvall. The MapEquation software package, 2013. URL http://-www.mapequation.org.

Paul J Fields, G Bruce Schaalje, and Matthew Roper. Examining a Misapplication of Nearest Shrucnken Centroid Classification to Investigate Book of Mormon Authorship. *The FARMS Review*, 23 (1): 87–111, 2011.

Santo Fortunato. Community detection in graphs. *Physics Reports*, 486 (3–5): 75 – 174, 2010. ISSN 0370-1573. doi: http://dx.doi.org/10.1016/j.physrep.2009.11.002. URL http://-www.sciencedirect.com/science/article/pii/S0370157309002841.

Santo Fortunato and Marc Barthelemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104 (1): 36–41, 2007.

Małgorzata Hantz. The Possible World of Terry Pratchett's Fantasy Fiction in Translation. Master's thesis, Institute of English Philology, Jagiellonian University, Kraków, 2013.

R. Heuser and L. Le-Khac. A Quantitative Literary History of 2,958 Nineteenth-Century British Novels: The Semantic Cohort Method, 2012. URL http://litlab.stanford.edu/-LiteraryLabPamphlet4.pdf.

David L. Hoover. Delta Prime? *Literary and Linguistic Computing*, 19 (4): 477–495, 2004a. doi: 10.1093/llc/19.4.477. URL http://llc.oxfordjournals.org/content/19/4/477.abstract.

David L. Hoover. Testing Burrows's Delta. *Literary and Linguistic Computing*, 19 (4): 453–475, 2004b. doi: 10.1093/llc/19.4.453. URL http://llc.oxfordjournals.org/content/19/4/-453.abstract.

David L. Hoover. Quantitative Analysis and Literary Studies. In Susan Schreibman and Ray Siemens, editors, *A Companion to Digital Literary Studies*. Blackwell, Oxford, 2008. URL http://www.digitalhumanities.org/companionDLS/.

Nancy Ide. Preparation and Analysis of Linguistic Corpora. In Susan Schreibman, Ray Siemens, and John Unsworth, editors, *A Companion to Digital Humanities*. Blackwell, Oxford, 2004. URL http://www.digitalhumanities.org/companion/.

Kim Luyckx and Walter Daelemans. The effect of author set size and data size in authorship attribution. *Literary and Linguistic Computing*, 26 (1): 35–55, 2011. doi: 10.1093/llc/fqq013. URL http://llc.oxfordjournals.org/content/26/1/35.abstract.

Willard McCarty. Knowing …: Modeling in Literary Studies. In Susan Schreibman and Ray Siemens, editors, *A Companion to Digital Literary Studies*. Blackwell, Oxford, 2008. URL http://www.digitalhumanities.org/companionDLS/.

Slav Petrov, Dipanjan Das, and Ryan McDonald. A Universal Part-of-Speech Tagset. In Nicoletta Calzolari, et al., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2089–2096, Istanbul, Turkey, 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7. URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/-274_Paper.pdf. ACL Anthology Identifier: L12-1115.

Maciej Piasecki. Polish Tagger TaKIPI: Rule Based Construction and Optimisation. *Task Quarterly*, 11 (1–2): 151–167, 2007.

Stephen Ramsay. Algorithmic Criticism. In Susan Schreibman and Ray Siemens, editors, *A Companion to Digital Literary Studies*. Blackwell, Oxford, 2008. URL http://-www.digitalhumanities.org/companionDLS/.

Jan Rybicki. Translation and Delta revisited: when we read translations, is it the author or the translator that we really read? In *Digital Humanities 2009: Conference Abstracts*,

pages 245–247, College Park (MA), 2009. University of Maryland. URL http://-www.mith2.umd.edu/dh09/wp-content/uploads/-dh09_conferencepreceedings_final.pdf.

Jan Rybicki. The Translator's Wife's traces. Alma Cardell Curtin and Jeremiah Curtin. *Przekładaniec. A Journal of Literary Translation*, 24: 89–109, 2010. doi: doi:10.4467/16891864ePC.12.005.0567.

Jan Rybicki. Alma Cardell Curtin and Jeremiah Curtin: the translator's wife's stylistic fingerprint. In *Digital Humanities 2011: Conference Abstracts*, pages 308–311, Stanford, 2011. Stanford University. URL http://dh2011abstracts.stanford.edu/xtf/view?docId=tei/ab-195.xml.

Jan Rybicki. The great mystery of the (almost) invisible translator: stylometry in translation. In M. Oakes and M. Ji, editors, *Quantitative Methods in Corpus-Based Translation Studies: A practical guide to descriptive translation research*, pages 231–248. John Benjamins, Amsterdam, 2012. doi: doi: 10.1075/scl.51.09ryb.

Jan Rybicki. Stylometryczna niewidzialność tłumacza. *Przekładaniec*, 27: 61–87, 2013. doi: doi:10.4467/16891864PC.13.004.1286.

Jan Rybicki and Maciej Eder. Deeper Delta across genres and languages: do we really need the most frequent words? *Literary and Linguistic Computing*, 2011. doi: 10.1093/llc/fqr031. URL http://llc.oxfordjournals.org/content/early/2011/07/14/-llc.fqr031.abstract.

Jan Rybicki and Magda Heydel. The stylistics and stylometry of collaborative translation: Woolf's Night and Day in Polish. *Literary and Linguistic Computing*, 28 (4): 708–717, 2013. doi: 10.1093/llc/fqt027. URL http://llc.oxfordjournals.org/content/28/4/708.abstract.

Beatrice Santorini. Part-of-speech tagging guidelines for the Penn Treebank Project (3rd revision). *Technical Reports (CIS)*, 1990. URL http://repository.upenn.edu/cis_reports/570. Paper 570.

Peter WH Smith and W Aldridge. Improving Authorship Attribution: Optimizing Burrows' Delta Method*. *Journal of Quantitative Linguistics*, 18 (1): 63–88, 2011.

Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60 (3): 538–556, 2009.

Kristina Toutanova and Christopher D Manning. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora (EMNLP/VLC-2000)*, pages 63–70. Association for Computational Linguistics, 2000.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL 2003*, pages 252–259. Association for Computational Linguistics, 2003.

V.A. Traag, P. Van Dooren, and Y. Nesterov. Narrow scope for resolution-free community detection. *arXiv:*, (1104.3083), 2011. [physics.soc-ph].

Brian Vickers. Shakespeare and authorship studies in the twenty-first century. *Shakespeare Quarterly*, 62 (1): 106–142, 2011.

Wikimedia Foundation. *Wikipedia: The Free Encyclopedia.* URL http://en.wikipedia.org.

Wikimedia Foundation. Świat Dysku. *Wikipedia: The Free Encyclopedia.* URL http://-pl.wikipedia.org/wiki/%C5%9Awiat_Dysku. [accessed: 10 September 2014]

William Winder. Writing Machines. In Susan Schreibman and Ray Siemens, editors, *A Companion to Digital Literary Studies*. Blackwell, Oxford, 2008. URL http://-

www.digitalhumanities.org/companionDLS/.

Marcin Woliński. System znaczników morfosyntaktycznych w korpusie IPI PAN. *Polonica*, XXII-XXIII: 39–55, 2003. URL http://nlp.ipipan.waw.pl/CORPUS/znakowanie.pdf.

Appendix A: Corpora

In this appendix I provide the list of texts used in the research, together with a short description. The English and Polish 100-novel corpora were compiled so as to retain comparable time span of creation of these works, as well as to provide a balance between male and female authors; the other corpora are biased.

The corpora are, in fact, collections of raw texts, without any annotations as one could expect [see Ide (2004) for an overview of corpus construction]. The annotation of parts of speech with available PoS taggers was a part of my (or my PC's) work, as described in **Appendices B.1-B.3**. At this point one should bear in mind that "perhaps the noise introduced by the NLP tools in the process of [syntactic or semantic feature] extraction is the crucial factor for their failure" (Stamatatos 2009).

There are some rules of what an ideal evaluation corpus should look like, which are listed by Stamatatos (2009): text length of training and test texts (in this case, unlike in Stamatatos's, the longer the better; Delta deals with long texts well; a text, however, should be fairly homogeneous in terms of style, topic, etc., which might not be viable for novels), controlling for genre, topic (in this case, genre is controlled for; the topic, if at all, is not controlled for directly), and other factors (age, education level, nationality, period, etc.), a well-defined set of candidate authors (not too small a group), distribution of the training corpus over the authors (the accuracy of authorship attribution might depend on the method; for Delta, the problem is rather the distribution of test corpus, since it influences the variances of word frequencies), several languages (because the style markers may vary according to the language). For developing my research further it is crucial to produce a range of benchmark corpora controlled for different characteristics, so that the stylistic markers can be singled out one by one.

The corpora in **Sections A.1-A.4** were compiled by dr Jan Rybicki, and made available to me. **Section A.5** contains titles collected by Michał Strojek, another MA student of dr Rybicki's.

The data on names, titles, and publishing dates were collected from Wikipidia (web-pages corresponding to particular authors) and (Chawton House Library) (introductory notes to the corresponding books).

A.1. English benchmark corpus small

A small-scale corpus of 27 classic English novels published between 1740-1876 by 11 authors, 1-3 books each, including 11 books written by female and 15 by male writers.

Table 1

File-id	Author	Title	Date[1]
ABrontë_Agnes	Brontë, Anne	Agnes Grey	1847
ABrontë_Tenant	Brontë, Anne	The Tenant of Wildfell Hall	1848
Austen_Emma	Austen, Jane	Emma	1815
Austen_Pride	Austen, Jane	Pride and Prejudice	1813
Austen_Sense	Austen, Jane	Sense and Sensibility	1811
CBrontë_Jane	Brontë, Charlotte	Jane Eyre	1847

[1] Date of publishing or, if known, of writing.

CBrontë_Professor	Brontë, Charlotte	The Professor	1857
CBrontë_Villette	Brontë, Charlotte	Villete	1853
Dickens_Bleak	Dickens, Charles	Bleak House	1853
Dickens_David	Dickens, Charles	David Copperfield	1849-1850
Dickens_Hard	Dickens, Charles	Hard Times: For These Times	1854
EBrontë_Wuthering	Brontë, Emily	Wuthering Heights	1845-1846
Eliot_Adam	Eliot, George	Adam Bede	1859
Eliot_Middlemarch	Eliot, George	Middlemarch	1872
Eliot_Mill	Eliot, George	The Mill on the Floss	1860
Fielding_Joseph	Fielding, Henry	The History of the Adventures of Joseph Andrews and of his Friend Mr. Abraham Adams	1742
Fielding_Tom	Fielding, Henry	The History of Tom Jones, a Foundling	1749
Richardson_Clarissa	Richardson, Samuel	Clarissa, or, the History of a Young Lady	1748
Richardson_Pamela	Richardson, Samuel	Pamela, or Virtue Rewarded	1740
Sterne_Sentimental	Sterne, Laurence	A Sentimental Journey Through France and Italy	1768
Sterne_Tristram	Sterne, Laurence	The Life and Opinions of Tristram Shandy, Gentleman	1759-1767
Thackeray_Barry	Thackeray, William Makepeace	Barry Lyndon	1844
Thackeray_ Pendennis	Thackeray, William Makepeace	The History of Pendennis	1850
Thackeray_Vanity	Thackeray, William Makepeace	Vanity Fair	1848
Trollope_Barchester	Trollope, Anthony	Barchester Towers	1857
Trollope_Phineas	Trollope, Anthony	Phineas Finn	1869
Trollope_Prime	Trollope, Anthony	The Prime Minister	1876

A.2. English benchmark corpus 100

A small-scale corpus of 100 English novels published between 1838-1937 by 33 authors, 3 books each (one authoring 4 books), including 33 books written by female and 67 by male writers.

Table 2

File-id	Author	Title	Date[2]
barclay_rosary	Barclay, Florence Louisa	The Rosary	1909
barclay_ladies	Barclay, Florence Louisa	The White Ladies of Worcester	1917
barclay_postern	Barclay, Florence Louisa	Through The Postern Gate	1911
bennet_helen	Bennett, Arnold	Helen With A High Hand	1910
bennet_imperial	Bennett, Arnold	Imperial Palace	1930
bennet_babylon	Bennett, Arnold	The Grand Babylon Hotel	1902
anon_clara	Blackmore, Richard Doddridge	Clara Vaughan	1853
blackmore_erema	Blackmore, Richard Doddridge	Erema	1876
blackmore_lorna	Blackmore, Richard Doddridge	Lorna Doone	1869
blackmore_ springhaven	Blackmore, Richard Doddridge	Springhaven	1887
braddon_quest	Braddon, Mary Elizabeth	Fenton's Quest	1871
braddon_audley	Braddon, Mary Elizabeth	Lady Audley's Secret	1862
braddon_fortune	Braddon, Mary, Elizabeth	Phantom Fortune	1883
cBrontë_jane	Brontë Charlotte	Jane Eyre	1847
cBrontë_shirley	Brontë Charlotte	Shirley	1849
cBrontë_villette	Brontë Charlotte	Villette	1853
lytton_kenelm	Bulwer-Lytton, Edward	Kenelm Chillingly	1873
lytton_novel	Bulwer-Lytton, Edward	My Novel	1853
lytton_what	Bulwer-Lytton, Edward	What Will He Do With It?	1858
burnett_princess	Burnett, Frances Hodgson	A Little Princess	1888
burnett_lord	Burnett, Frances Hodgson	Little Lord Fauntleroy	1885

[2] Date of publishing or, if known, of writing.

burnett_garden	Burnett, Frances Hodgson	The Secret Garden	1910
chesterton_ innocence	Chesterton, Gilbert Keith	The Innocence of Father Brown	1911
chesterton_ thursday	Chesterton, Gilbert Keith	The Man Who Was Thursday	1908
chesterton_ napoleon	Chesterton, Gilbert Keith	The Napoleon of Notting Hill	1904
collins_basil	Collins, Wilkie	Basil	1852
collins_cain	Collins, Wilkie	The Legacy of Cain	1889
collins_woman	Collins, Wilkie	The Woman in White	1860
conrad_almayer	Conrad, Joseph	Almayer's Folly	1895
conrad_nostromo	Conrad, Joseph	Nostromo	1904
conrad_rover	Conrad, Joseph	The Rover	1922
corelli_romance	Corelli, Marie	A Romance of Two Worlds	1886
corelli_innocent	Corelli, Marie	Innocent	1914
corelli_satan	Corelli, Marie	The Sorrows of Satan	1895
dickens_bleak	Dickens, Charles	Bleak House	1853
dickens_ expectations	Dickens, Charles	Great Expectations	1861
dickens_oliver	Dickens, Charles	Oliver Twist	1838
doyle_micah	Doyle, Arthur Conan	Micah Clarke	1888
doyle_hound	Doyle, Arthur Conan	The Hound of the Baskervilles	1902
doyle_lost	Doyle, Arthur Conan	The Lost World	1912
eliot_adam	Eliot, George	Adam Bede	1859
eliot_daniel	Eliot, George	Daniel Deronda	1876
eliot_felix	Eliot, George	Felix Holt, the Radical	1866
ford_girl	Ford, Ford Madox	An English Girl	1907
ford_soldier	Ford, Ford Madox	The Good Soldier	1915
ford_post	Ford, Ford Madox	The Last Post	1928
forster_room	Forster, Edward Morgan	A Room with A View	1908
forster_howards	Forster, Edward Morgan	Howard's End	1910
forster_angels	Forster, Edward Morgan	Where Angels Fear to Tread	1905
galsworthy_river	Galsworthy, John	Over the River	1933
galsworthy_saints	Galsworthy, John	Saint's Progress	1919
galsworthy_man	Galsworthy, John	The Man of Property	1906
gaskell_ruth	Gaskell, Elizabeth	Ruth	1853
gaskell_lovers	Gaskell, Elizabeth	Sylvia's Lovers	1863
gaskell_wives	Gaskell, Elizabeth	Wives and Daughters	1865
gissing_women	Gissing, George	The Odd Women	1893
gissing_unclassed	Gissing, George	The Unclassed	1884

gissing_warburton	Gissing, George	Will Warburton	1905
hardy_madding	Hardy, Thomas	Far from the Madding Crowd	1874
hardy_jude	Hardy, Thomas	Jude the Obscure	1895
hardy_tess	Hardy, Thomas	Tess of the d'Urbervilles	1891
james_hudson	James, Henry	Roderick Hudson	1875
james_ambassadors	James, Henry	The Ambassadors	1903
james_muse	James, Henry	The Tragic Muse	1890
kipling_captains	Kipling, Rudyard	Captains Courageous	1897
kipling_kim	Kipling, Rudyard	Kim	1901
kipling_light	Kipling, Rudyard	The Light That Failed	1890
lawrence_serpent	Lawrence, David Herbert	The Plumed Serpent	1926
lawrence_peacock	Lawrence, David Herbert	The White Peacock	1911
lawrence_women	Lawrence, David Herbert	Women in Love	1920
lee_brown	Lee, Vernon	Miss Brown	1884
lee_penelope	Lee, Vernon	Penelope Brandling	1903
lee_albany	Lee, Vernon	The Countess of Albany	1884
meredith_richmond	Meredith, George	The Adventures of Harry Richmond	1871
meredith_marriage	Meredith, George	The Amazing Marriage	1895
meredith_feverel	Meredith, George	The Ordeal of Richard Feverel	1859
morris_roots	Morris, William	The Roots of the Mountains	1890
morris_water	Morris, William	The Water of the Wondrous Isles	1897
morris_wood	Morris, William	The Wood Beyond the World	1894
haggard_mines	Rider, Haggard Henry	King Solomon's Mines	1885
haggard_sheallan	Rider, Haggard Henry	She and Allan	1921
haggard_mist	Rider, Haggard Henry	The People of the Mist	1894
schreiner_african	Schreiner, Olive	The Story of an African Farm	1883
schreiner_trooper	Schreiner, Olive	Trooper Peter Halket of Mashonaland	1897
schreiner_undine	Schreiner, Olive	Undine	1929
stevenson_catriona	Stevenson, Robert Louis	Catriona	1893
stevenson_arrow	Stevenson, Robert Louis	The Black Arrow	1883
stevenson_island	Stevenson, Robert Louis	Treasure Island	1882

thackeray_esmond	Thackeray, William Makepeace	The History of Henry Esmond	1852
thackeray_ pendennis	Thackeray, William Makepeace	The History of Pendennis	1850
thackeray_virginians	Thackeray, William Makepeace	The Virginians	1859
trollope_angel	Trollope, Anthony	Ayala's Angel	1878
trollope_phineas	Trollope, Anthony	Phineas Finn	1869
trollope_warden	Trollope, Anthony	The Warden	1855
ward_harvest	Ward, Mary Augusta	Harvest	1920
ward_milly	Ward, Mary Augusta	Milly and Olly	1881
ward_ashe	Ward, Mary Augusta	The Marriage of William Ashe	1905
woolf_night	Woolf, Virginia	Night and Day	1919
woolf_years	Woolf, Virginia	The Years	1937
woolf_lighthouse	Woolf, Virginia	To the Lighthouse	1927

A.3. English corpus 500

A medium-scale corpus of 499 English novels published between 1704-2013 by 140 authors or collaborations; the mean size of authorial groups is 3.56 (see **Figure 17** below for a histogram). The topics, subgenres, editorial policies, etc. vary, as could be expected from the time-scale alone.

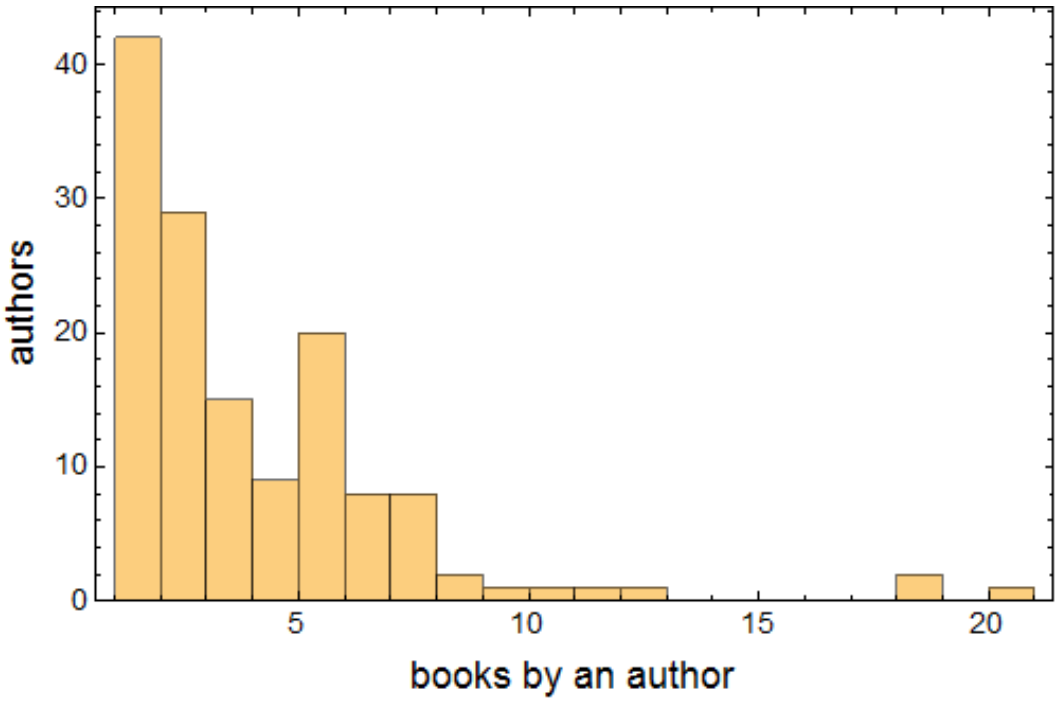

Figure 17 The histogram of the number of books by a given author for EN500 benchmark.

It should be added that I excluded anonymous texts from the corpus. In the process it also became apparent that two books included in the corpus as different (Aubin's *The Life of Charlotta Du Pont, an English Lady* and *The Inhuman Stepmother; or the History of Miss Harriot Montague*) are in fact the same novel, as indicated by Delta distance < 0.1; see the introduction to *Charlotta* (Chawton House Library, *Charlotta*).

Table 3

File-id	Author	Title	Date[3]
Freeman_Thorncase	-	-	-
Anderson_Marching	Anderson, Sherwood	Marching Men	1917
Anderson_White	Anderson, Sherwood	Poor White	1920
Anderson_Windy	Anderson, Sherwood	Windy McPherson's Son	1916
Anderson_ Winesburg	Anderson, Sherwood	Winesburg, Ohio	1919

[3] Date of publishing or, if known, of writing.

Aubin_Charlotta	Aubin, Penelope	The Life of Charlotta Du Pont, an English lady; taken from her own memoirs	1723
Austen_Emma	Austen, Jane	Emma	1815
Austen_Mansfield	Austen, Jane	Mansfield Park	1814
Austen_Northanger	Austen, Jane	Northanger Abbey	1818
Austen_Persuasion	Austen, Jane	Persuasion	1818
Austen_Pride	Austen, Jane	Pride and Prejudice	1813
Austen_Sense	Austen, Jane	Sense and Sensibility	1811
Beckett_Malone	Beckett, Samuel Barclay	Malone Dies	1956
Beckett_Molloy	Beckett, Samuel Barclay	Molloy	1955
Beckett_Unnamable	Beckett, Samuel Barclay	The Unnamable	1958
Beckford_Azemia	Beckford, William Thomas	Azemia	1797
Beckford_Vathek	Beckford, William Thomas	Vathek	1781
Bennett_Agnes	Bennett, Anna Maria	Agnes de-Courci: a Domestic Tale	1789
bennet_babylon	Bennett, Arnold	The Grand Babylon Hotel	1902
bennet_helen	Bennett, Arnold	Helen With A High Hand	1910
bennet_imperial	Bennett, Arnold	Imperial Palace	1930
Bentley_Trent	Bentley, Edmund Clerihew	?Trent's Last Case	1913
Bowen_Friends	Bowen, Elizabeth	Friends and Relations	1931
Bowen_Hotel	Bowen, Elizabeth	The Hotel	1927
Bradford_Plymouth	Bradford, William	Of Plymouth Plantation	1930, 1946--1950
ABrontë_Agnes	Brontë, Anne	Agnes Grey	1847
ABrontë_Tenant	Brontë, Anne	The Tenant of Wildfell Hall	1848
CBrontë_Jane	Brontë, Charlotte	Jane Eyre	1847
CBrontë_Professor	Brontë, Charlotte	The Professor	1857
cBrontë_shirley	Brontë, Charlotte	Shirley	1849
CBrontë_Villette	Brontë, Charlotte	Villete	1853
EBrontë_Wuthering	Brontë, Emily	Wuthering Heights	1845--1846
Brown_Angels	Brown, Dan	Angels & Demons	2000
Brown_Davinci	Brown, Dan	The Da Vinci Code	2003
Brown_Point	Brown, Dan	Deception Point	2001
Lytton_Barons	Bulwer-Lytton, Edward	The Last of the Barons	1843
Lytton_Harold	Bulwer-Lytton, Edward	Harold, the Last of the Saxons	1848
Lytton_Pompeii	Bulwer-Lytton,	The Last Days of Pompeii	1834

	Edward		
Lytton_Rienzi	Bulwer-Lytton, Edward	Rienzi, the last of the Roman tribunes	1835
Lytton_Zanoni	Bulwer-Lytton, Edward	Zanoni	1842
Burney_Camilla	Burney, Frances	Camilla: Or, A Picture of Youth	1796
Burney_Cecilia	Burney, Frances	Cecilia: Or, Memoirs of an Heiress	1782
Burney_Darblay	Burney, Frances	The Diary and Letters of Madame D'Arblay	1904
Burney_Evelina	Burney, Frances	Evelina: Or The History of A Young Lady's Entrance into the World	1778
Burney_Wanderer	Burney, Frances	The Wanderer: Or, Female Difficulties	1814
Canning_Offspring	Canning, J.A.	Offspring	2013
Capote_Voices	Capote, Truman	Other Voices, Other Rooms	1948
Carver_OldWoman	Carver, Mrs	The Old Woman	1800
Cary_Johnson	Cary, Joyce	Mister Johnson	1939
Cather_Antonia	Cather, Willa Sibert	My Ántonia	1918
Cather_Bridge	Cather, Willa Sibert	Alexander's Bridge	1912
Cather_Lark	Cather, Willa Sibert	The Song of the Lark	1915
Cather_Ours	Cather, Willa Sibert	One of Ours	1922
Cather_Pioneers	Cather, Willa Sibert	O Pioneers!	1913
Catton_Luminaries	Catton, Eleanor	The Luminaries	2013
Charlton_Parisian	Charlton, Mary	The Parisian; or, Genunine Anecdotes of Distinguished and Noble Characters	1794
chesterton_innocence	Chesterton, Gilbert Keith	The Innocence of Father Brown	1911
Chesterton_Scandal	Chesterton, Gilbert Keith	The Scandal of Father Brown	1935
Chesterton_Secret	Chesterton, Gilbert Keith	The Secret of Father Brown	1927
chesterton_thursday	Chesterton, Gilbert Keith	The Man Who Was Thursday	1908
Chesterton_Wisdom	Chesterton, Gilbert Keith	The Wit and Wisdom of GK Chesterton.	1911
Christie_Abc	Christie, Agatha	The A.B.C. Murders	1936
Christie_Ackroyd	Christie, Agatha	The Murder of Roger Ackroyd	1926
Christie_AndThen	Christie, Agatha	Ten Little Niggers	1939
Christie_Appointment	Christie, Agatha	Appointment with Death	1938
Christie_Cards	Christie, Agatha	Cards on the Table	1936
Christie_Chimneys	Christie, Agatha	The Secret of Chimneys	1925

Christie_Curtain	Christie, Agatha	Curtain	1975
Christie_Orient_Express	Christie, Agatha	Murder on the Orient Express	1934
Clancy_Games	Clancy, Tom	Patriot Games	1987
Clancy_Redoctober	Clancy, Tom	The Hunt for Red October	1984
Coben_Breaker	Coben, Harlan	Deal Breaker	1995
Coben_Dropshot	Coben, Harlan	Drop Shot	1996
Coben_Fade	Coben, Harlan	Fade Away	1996
Coben_Falsemove	Coben, Harlan	One False Move	1998
Coben_Gone	Coben, Harlan	Gone for Good	2002
Coben_NoSecond	Coben, Harlan	No Second Chance	2003
Coben_Tell	Coben, Harlan	Tell No One	2001
WCollins_Armadale	Collins, Wilkie	Armadale	1866
WCollins_Hotel	Collins, Wilkie	The Haunted Hotel	1878
WCollins_Moonstone	Collins, Wilkie	The Moonstone	1868
WCollins_Robe	Collins, Wilkie	The Black Robe	1881
WCollins_Woman	Collins, Wilkie	The Woman in White	1860
Compton_Dolores	Compton-Burnett, Dame Ivy	Dolores	1911
Compton_Family	Compton-Burnett, Dame Ivy	A Family and a Fortune	1939
Compton_Men	Compton-Burnett, Dame Ivy	Men and Wives	1931
Conrad_Agent	Conrad, Joseph	The Secret Agent	1907
conrad_almayer	Conrad, Joseph	Almayer's Folly	1895
Conrad_ArrowGld	Conrad, Joseph	The Arrow of Gold	1919
Conrad_Chance	Conrad, Joseph	Chance	1913
Conrad_Duel	Conrad, Joseph	The Duel: A Military Story	1908
Conrad_EndTethr	Conrad, Joseph	The End of the Tether	1902
Conrad_Falk	Conrad, Joseph	Falk	1901
Conrad_Freya	Conrad, Joseph	Freya of the Seven Isles	1910--1911
Conrad_Heart	Conrad, Joseph	Heart of Darkness	1899
Conrad_Lord	Conrad, Joseph	Lord Jim	1900
Conrad_NiggerN	Conrad, Joseph	The Nigger of the 'Narcissus'	1897
conrad_nostromo	Conrad, Joseph	Nostromo	1904
Conrad_OutcastI	Conrad, Joseph	An Outcast of the Islands	1896
Conrad_Rescue	Conrad, Joseph	The Rescue	1920
conrad_rover	Conrad, Joseph	The Rover	1922
Conrad_ShaddowL	Conrad, Joseph	The Shadow Line	1917
Conrad_ SmileFortune	Conrad, Joseph	A Smile of Fortune	1910
Conrad_Typhoon	Conrad, Joseph	Typhoon	1899--1902
Conrad_Victory	Conrad, Joseph	Victory	1915

Conrad_Western	Conrad, Joseph	Under Western Eyes	1911
ConFord_Inheritors	Conrad, Joseph and Ford, Ford Madox	The Inheritors	1901
ConFord_Nature	Conrad, Joseph and Ford, Ford Madox	The Nature of Crime	1923
ConFord_Romance	Conrad, Joseph and Ford, Ford Madox	Romance	1903
Cooper_CarolineHerbert	Cooper, Maria Susanna	The Wife; or, Caroline Herbert	1813
Craik_Stella	Craik, Helen	Stella of the North, or the	1802
Defoe_Cruzoe	Defoe, Daniel	Robinson Crusoe	1719
Defoe_Moll	Defoe, Daniel	Moll Flanders	1722
Defoe_Roxana	Defoe, Daniel	Roxana: The Fortunate Mistress	1724
Defoe_Singleton	Defoe, Daniel	Captain Singleton	1720
Dickens_Barnaby	Dickens, Charles	Barnaby Rudge: A Tale of the Riots of 'Eighty	1841
Dickens_Bleak	Dickens, Charles	Bleak House	1853
Dickens_Boz	Dickens, Charles	Sketches by Boz	1836
Dickens_Chimes	Dickens, Charles	The Chimes	1844
Dickens_Christmas	Dickens, Charles	A Christmas Carol	1843
Dickens_Cities	Dickens, Charles	A Tale of Two Cities	1859
Dickens_Cricket	Dickens, Charles	The Cricket on the Hearth	1845
Dickens_Curiosity	Dickens, Charles	The Old Curiosity Shop	1840--1841
Dickens_David	Dickens, Charles	David Copperfield	1849--1850
Dickens_Dombey	Dickens, Charles	Dombey and Son	1846--1848
Dickens_Dorrit	Dickens, Charles	Little Dorrit	1855--1857
Dickens_Edwin	Dickens, Charles	The Mystery of Edwin Drood	1870
dickens_ expectations	Dickens, Charles	Great Expectations	1861
Dickens_Hard	Dickens, Charles	Hard Times: For These Times	1854
Dickens_Mutual	Dickens, Charles	Our Mutual Friend	1864--1865
Dickens_Nicholas	Dickens, Charles	The Life and Adventures of Nicholas Nickleby	1838--1839
dickens_oliver	Dickens, Charles	Oliver Twist	1838
Dickens_Pickwick	Dickens, Charles	The Posthumous Papers of the Pickwick Club	1836--1837
Disraeli_Coningsby	Disraeli, Benjamin	Coningsby, or the New Generation	1844
Disraeli_Endymion	Disraeli, Benjamin	Endymion	1880

Disraeli_Lothair	Disraeli, Benjamin	Lothair	1870
Disraeli_Sybil	Disraeli, Benjamin	Sybil, or The Two Nations	1845
Disraeli_Vivian	Disraeli, Benjamin	Vivian Grey	1826
Passos_Initiation	Dos Passos, John Roderigo	One Man's Initiation: 1917	1917 1920
Passos_Soldiers	Dos Passos, John Roderigo	Three Soldiers	1921
Doyle_Adventures	Doyle, Arthur Conan	The Adventures of Sherlock Holmes	1892
Doyle_Lastbow	Doyle, Arthur Conan	His Last Bow	1917
Doyle_MaracotDeep	Doyle, Arthur Conan	The Maracot Deep	1929
Doyle_Study	Doyle, Arthur Conan	A Study in Scarlet	1887
Doyle_TheHound	Doyle, Arthur Conan	The Hound of the Baskervilles	1902
Doyle_TheLostWorld	Doyle, Arthur Conan	The Lost World	1912
Edgeworth_Absentee	Edgeworth, Maria	The Absentee	1812
Edgeworth_Assistant	Edgeworth, Maria	The Parent's Assistant	1796
Edgeworth_Belinda	Edgeworth, Maria	Belinda	1801
Edgeworth_Ennui	Edgeworth, Maria	Ennui	1809
Edgeworth_Forester	Edgeworth, Maria	Forester	1872
Edgeworth_Rackrent	Edgeworth, Maria	Castle Rackrent	1800
Edgeworth_Vivian	Edgeworth, Maria	Vivian	1809--1812
eliot_adam	Eliot, George	Adam Bede	1859
eliot_daniel	Eliot, George	Daniel Deronda	1876
eliot_felix	Eliot, George	Felix Holt, the Radical	1866
Eliot_Middlemarch	Eliot, George	Middlemarch	1872
Eliot_Mill	Eliot, George	The Mill on the Floss	1860
Eliot_Romola	Eliot, George	Romola	1863
Eliot_Silas	Eliot, George	Silas Marner,	1861
Ellison_Invisible	Ellison, Ralph	Invisible Man	1952
Faulkner_Absalom	Faulkner, William	Absalom, Absalom!	1936
Faulkner_Dying	Faulkner, William	As I Lay Dying	1930
Faulkner_Light	Faulkner, William	Light in August	1932
Faulkner_Moses	Faulkner, William	Go Down, Moses	1941
Faulkner_Sound	Faulkner, William	The Sound and the Fury	1929
Fielding_Amelia	Faulkner, William	Amelia	1751
Fielding_Joseph	Fielding, Henry	The History of the Adventures of Joseph Andrews and of his Friend	1742

		Mr. Abraham Adams	
Fielding_Shamela	Fielding, Henry	An Apology for the Life of Mrs. Shamela Andrews	1741
Fielding_Tom	Fielding, Henry	The History of Tom Jones, a Foundling	1749
Fitzgerald_Beautiful	Fitzgerald, Francis Scott	The Beautiful and Damned	1922
Fitzgerald_Gatsby	Fitzgerald, Francis Scott	The Great Gatsby	1925
Fitzgerald_Paradise	Fitzgerald, Francis Scott	This Side of Paradise	1920
Fitzgerald_Tender	Fitzgerald, Francis Scott	Tender Is the Night	1934
forster_angels	Forster, Edward Morgan	Where Angels Fear to Tread	1905
forster_howards	Forster, Edward Morgan	Howard's End	1910
Forster_Journey	Forster, Edward Morgan	The Longest Journey	1907
Forster_Maurice	Forster, Edward Morgan	Maurice	1913--1914
Forster_Passage	Forster, Edward Morgan	A Passage to India	1924
Forster_View	Forster, Edward Morgan	A Room with a View	1908
Foster_Corinna	Forster, Edward Morgan	The Corinna of England, and a Heroine in the Shade: A Modern Romance	1809
Foster_Substance	Forster, Edward Morgan	Substance and shadow, or, The fisherman's daughters of Brighton : a patchwork story	1812
Freeman_Bone	Freeman, Brian	The Bone House	2011
Galsworthy_Chancery	Galsworthy, John	In Chancery	1920
Galsworthy_Let	Galsworthy, John	To Let	1921
galsworthy_man	Galsworthy, John	The Man of Property	1906
Galsworthy_Monkey	Galsworthy, John	The White Monkey	1924
Galsworthy_Pharisees	Galsworthy, John	The Island Pharisees	1904
Gaskell_Barton	Gaskell, Elizabeth	Mary Barton	1848
Gaskell_Lovers	Gaskell, Elizabeth	Sylvia's Lovers	1863
Gaskell_NorthSouth	Gaskell, Elizabeth	North and South	1854--1855
Gaskell_Ruth	Gaskell, Elizabeth	Ruth	1853
Gaskell_Wives	Gaskell, Elizabeth	Wives and Daughters: An Everyday Story	1865

Glasgow_Veiniron	Glasgow, Ellen	Vein of Iron	1935
Godwin_Caleb	Godwin, William	Things as They Are; or, The Adventures of Caleb Williams	1794
Godwin_Imogen	Godwin, William	Imogen: A Pastoral Romance	-
Golding_Inheritors	Golding, William	The Inheritors	1955
Golding_Lord	Golding, William	Lord of the Flies	1954
Golding_Rites	Golding, William	Rites of Passage	1980
Golding_Spire	Golding, William	The Spire	1964
Goldsmith_Vicar	Goldsmith, Oliver	The Vicar of Wakefield: A Tale. Supposed to be written by Himself	1761-1762
Green_Loving	Green, Henry	Loving	1945
Green_Party	Green, Henry	Party Going	1939
Green_Romance	Green, Henry	-	-
Greene_Brighton	Greene, Graham	Brighton Rock	1938
Greene_BurntOut	Greene, Graham	A Burnt-Out Case	1960
Greene_Confidential	Greene, Graham	The Confidential Agent	1939
Greene_Havana	Greene, Graham	Our Man in Havana	1958
Greene_Ministry	Greene, Graham	The Ministry of Fear	1943
Greene_Power	Greene, Graham	The Power and the Glory	1940
Grisham_Broker	Grisham, John	The Broker	2005
Grisham_Christmas	Grisham, John	Skipping Christmas†	2001
Grisham_Partner	Grisham, John	The Partner	1997
Grisham_Pelican	Grisham, John	The Pelican Brief	1992
Grisham_StreetLawyer	Grisham, John	The Street Lawyer	1998
Hall_Lamp	Hall, Radclyffe	The Unlit Lamp	1924
Hall_Well	Hall, Radclyffe	The Well of Loneliness	1928
Hardy_BlueEyes	Hardy, Thomas	A Pair of Blue Eyes: A Novel	1873
Hardy_Greenwood	Hardy, Thomas	Under the Greenwood Tree: A Rural Painting of the Dutch School	1872
hardy_jude	Hardy, Thomas	Jude the Obscure	1895
hardy_madding	Hardy, Thomas	Far from the Madding Crowd	1874
Hardy_Native	Hardy, Thomas	The Return of the Native	1878
hardy_tess	Hardy, Thomas	Tess of the d'Urbervilles	1891
Hardy_Woodlanders	Hardy, Thomas	The Woodlanders	1887
Harvey_Anything	Harvey, Jane	-	-
Harvey_Tynemouth	Harvey, Jane	The Castle of Tynemouth	1806
Hatton_Lovers	Hatton, Ann	Lovers and Friends; or, Modern Attachments	1821
Hawthorne_Blithedale	Hawthorne, Nathaniel	The Blithedale Romance	1852
Hawthorne_Fansha	Hawthorne,	Fanshawe	1828

we	Nathaniel		
Hawthorne_Marble Faun	Hawthorne, Nathaniel	The Marble Faun: Or, The Romance of Monte Beni	1860
Hawthorne_Scarlet	Hawthorne, Nathaniel	The Scarlet Letter	1850
Hawthorne_SevenGables	Hawthorne, Nathaniel	The House of the Seven Gables	1851
Helme_Magdalen	Helme, Elizabeth	Magdalen: or, The penitant of Godstow. An historical novel	1813
Hemingway_Across	Hemingway, Ernest Miller	Across the River and into the Trees	1950
Hemingway_Bell	Hemingway, Ernest Miller	For Whom the Bell Tolls	1940
Hemingway_Farewell	Hemingway, Ernest Miller	A Farewell to Arms	1929
Hemingway_Fiesta	Hemingway, Ernest Miller	The Sun Also Rises	1926
Hemingway_Menwithout	Hemingway, Ernest Miller	Men Without Women	1927
Hemingway_Oldman	Hemingway, Ernest Miller	The Old Man and the Sea	1951
james_ambassadors	Henry, James	The Ambassadors	1903
James_American	Henry, James	The American	1876--1877
James_Awkward	Henry, James	the awkward age	1899
James_Bostonians	Henry, James	The Bostonians, a novel	1885-1885
James_Confidence	Henry, James	Confidence	1879
James_Europeans	Henry, James	The Europeans	1878
James_Golden	Henry, James	The golden bowl	1904
James_Maisie	Henry, James	What Maisie knew	1897
James_Portrait	Henry, James	The portrait of a lady	1880--1881
James_Poynton	Henry, James	The spoils of Poynton	1896
James_Princess	Henry, James	he Princess Casamassima	1886
James_Reverbera	Henry, James	The Reverberator	1888
james_hudson	Henry, James	Roderick Hudson	1875
James_Sacred	Henry, James	The Sacred Fount	1901
james_muse	Henry, James	The Tragic Muse	1890
James_Washington	Henry, James	Washington Square	1880
James_Watch	Henry, James	Watch and Ward	1871
James_Wings	Henry, James	The Wings of the Dove	1902
Lawrence_Aarons	Herbert, Lawrence David	Aaron's Rod	1922
Lawrence_ Chatterleys	Herbert, Lawrence David	Lady Chatterley's Lover	1928

lawrence_peacock	Herbert, Lawrence David	The White Peacock	1911
Lawrence_Rainbow	Herbert, Lawrence David	The Rainbow	1915
Lawrence_Sons	Herbert, Lawrence David	Sons and Lovers	1913
lawrence_women	Herbert, Lawrence David	Women in Love	1920
Hughes_Caroline	Hughes, Anne	Caroline; or, the Diversities of Fortune	1787
Hunter_Unexpected	Hunter, Rachel	The Unexpected Legacy	1804
Huxley_Brave	Huxley, Aldous Leonard	Brave New World	1932
Huxley_Crome	Huxley, Aldous Leonard	Crome Yellow	1921
Huxley_Gaza	Huxley, Aldous Leonard	Eyeless in Gaza	1936
Huxley_Mortal	Huxley, Aldous Leonard	Mortal Coils	1922
Huxley_Point	Huxley, Aldous Leonard	Point Counter Point	1928
Jacson_Isabella	Jacson, Frances	Isabella. A Novel	1823
Jacson_Things	Jacson, Frances	Things by their Right Names	1812
Johnson_Francis	Johnson, Mrs.	Francis, the Philanthropist: an unfashionable tale	1786
Sjohnson_Rasselas	Johnson, Samuel	The History of Rasselas, Prince of Abissinia	1759
Sjohnson_Scotland	Johnson, Samuel	A Journey to the Western Islands of Scotland	1775
Joyce_Dubliners	Joyce, James	Dubliners	1914
Joyce_Finnegans	Joyce, James	Finnegans Wake	1939
Joyce_Portrait	Joyce, James	A Portrait of the Artist as a Young Man	1916
Joyce_Ulysses	Joyce, James	Ulysses	1922
kipling_captains	Kipling, Rudyard	Captains Courageous	1897
Kipling_Jungle	Kipling, Rudyard	The Jungle Book	1894
kipling_kim	Kipling, Rudyard	Kim	1901
Kipling_Puck	Kipling, Rudyard	Puck of Pook's Hill	1906
Kipling_Rewards	Kipling, Rudyard	Rewards and Fairies	1910
Koontz_Brotherodd	Koontz, Dean	Brother Odd	2006
Koontz_December	Koontz, Dean	The Door To December	1985
Koontz_DoorFromHeaven	Koontz, Dean	One Door Away from Heaven	2001
Koontz_FearNothing	Koontz, Dean	Fear Nothing	1998
Koontz_Hideaway	Koontz, Dean	Hideaway	1992
Sterne_Sentimental	Laurence, Sterne	A Sentimental Journey Through France and Italy	1768

Sterne_Tristram	Laurence, Sterne	The Life and Opinions of Tristram Shandy, Gentleman	1759-1767
Harperlee_Mockingbird	Lee, Harper	To Kill a Mockingbird	1960
Lessing_Ben	Lessing, Doris May	Ben, in the World	2000
Lessing_Child	Lessing, Doris May	The Fifth Child	1988
Lessing_City	Lessing, Doris May	The Four-Gated City	1969
Lessing_Dream	Lessing, Doris May	The Sweetest Dream	2001
Lessing_General	Lessing, Doris May	The Story of General Dann and Mara's Daughter, Griot and the Snow Dog	2005
Lessing_MaraDann	Lessing, Doris May	Mara and Dann	1999
Lessing_Orkney	Lessing, Doris May	The Temptation of Jack Orkney: Collected Stories, Vol. 2	1978
Lessing_Planet	Lessing, Doris May	The Making of the Representative for Planet 8	1982
Lessing_Sirian	Lessing, Doris May	The Sirian Experiments	1980
ALewis_Vicissitudes	Lewis, Alethea	Vicissitudes in Genteel Life	1794
Lewis_Battle	Lewis, Clive Staples	The Last Battle	1956
Lewis_Caspian	Lewis, Clive Staples	Prince Caspian	1951
Lewis_Chair	Lewis, Clive Staples	The Silver Chair	1953
Lewis_Horse	Lewis, Clive Staples	The Horse and His Boy	1954
Lewis_Lion	Lewis, Clive Staples	The Lion, the Witch and the Wardrobe	1950
Lewis_Nephew	Lewis, Clive Staples	The Magician's Nephew	1955
Lewis_Voyage	Lewis, Clive Staples	The Voyage of the Dawn Treader	1952
Mlewis_Bravo	Lewis, Matthew Gregory	The Bravo of Venice	1805
Mlewis_Monk	Lewis, Matthew Gregory	The Monk	1796
Wlewis_Condemned	Lewis, Wyndham	Self Condemned	1954
Wlewis_Tarr	Lewis, Wyndham	Tarr	-
Ludlum_Bidentity	Ludlum, Robert	The Bourne Identity	1980
Ludlum_BSupremacy	Ludlum, Robert	The Bourne Supremacy	1986
Ludlum_Halidon	Ludlum, Robert	The Cry of the Halidon	1974
Ludlum_Icarus	Ludlum, Robert	The Icarus Agenda	1988
MacDonald_Curdie	MacDonald, George	The Princess and Curdie	1883
MacDonald_Goblin	MacDonald, George	The Princess and the Goblin	1872
Mackenzie_Irish	Mackenzie, Anna Maria	The Irish Guardian, or, Errors of Eccentricity	1809
Mackenzie_Monmouth	Mackenzie, Anna Maria	Monmouth: a Tale, Founded on Historic Facts	1790

ford_girl	Madox, Ford Ford	An English Girl	1907
Ford_ARingForNancy	Madox, Ford Ford	Ring for Nancy : a sheer comedy	1913
Ford_Crowned	Madox, Ford Ford	The Fifth Queen Crowned	1908
Ford_Ladies	Madox, Ford Ford	Ladies Whose Bright Eyes	1911
Ford_MrApollo	Madox, Ford Ford	Mr Apollo	1908
Ford_PrivySeal	Madox, Ford Ford	Privy Seal	1907
Ford_Queen	Madox, Ford Ford	The Fifth Queen	1906
ford_soldier	Madox, Ford Ford	The Good Soldier	1915
Ford_TheBenefactor	Madox, Ford Ford	The Benefactor	1905
Ford_TheHalfMoon	Madox, Ford Ford	The Half Moon	1909
Mansfield_Garden	Mansfield, Katherine	The Garden Party: and Other Stories	1922
Mansfield_German	Mansfield, Katherine	In a German Pension	1911
Mansfield_Bliss	Mansfield, Katherine	Bliss: and Other Stories	1920
Martin_Enchantress	Martin, Mrs.	The Enchantress; or, Where Shall I Find Her? A Tale	1801
Mathews_Simple	Mathews, Mrs.	Simple Facts; or, the History of an Orphan	1793
Maugham_Bondage	Maugham, William Somerset	Of Human Bondage	1934
Maugham_Hero	Maugham, William Somerset	The Hero	1901
Maugham_Liza	Maugham, William Somerset	Liza of Lambeth	1897
Maugham_Magician	Maugham, William Somerset	The Magician	1926
Maugham_Moon	Maugham, William Somerset	The Moon and Sixpence	1919
McNeile_Black	McNeile, Herman Cyril	The Black Gang	1922
McNeile_Bulldog	McNeile, Herman Cyril	Bull-Dog Drummond	1920
Melville_Bartleby	Melville, Herman	Bartleby, the Scrivener	1853
Melville_Mobydick	Melville, Herman	Moby-Dick; or, The Whale	1851
Melville_Omoo	Melville, Herman	Omoo: A Narrative of Adventures in the South Seas	1847
Melville_Redburn	Melville, Herman	Redburn: His First Voyage	1849
Melville_Typee	Melville, Herman	Typee: A Peep at Polynesian Life	1846
Meredith_Diana	Meredith, George	Diana of the Crossways	1885
meredith_feverel	Meredith, George	The Ordeal of Richard Feverel	1859
meredith_richmond	Meredith, George	The Adventures of Harry	1871

		Richmond	
Mitchell_Gonewind	Mitchell, Margaret	Gone with the Wind	1936
Montgomery_Gables	Montgomery, Lucy Maud	Anne of Green Gables	1908
Montgomery_Ingleside	Montgomery, Lucy Maud	Anne of Ingleside	1939
Montgomery_Island	Montgomery, Lucy Maud	Anne of the Island	1915
Montgomery_Quoted	Montgomery, Lucy Maud	The Blythes Are Quoted	1942
Morris_Child	Morris, William	Child Christopher and Goldilind the Fair	1895
Morris_House	Morris, William	A Tale of the House of the Wolfings, and All the Kindreds of the Mark Written in Prose and in Verse	1889
Morris_JohnBall	Morris, William	A Dream of John Ball	1888
Morris_Plain	Morris, William	The Story of the Glittering Plain	1891
morris_roots	Morris, William	The Roots of the Mountains	1890
Morris_Signs	Morris, William	Signs of Change	1888
Morris_Story	Morris, William	The Story of Sigurd the Volsung and the Fall of the Niblungs	1877
Morris_Sundering	Morris, William	The Sundering Flood	1897
morris_water	Morris, William	The Water of the Wondrous Isles	1897
Morris_Well	Morris, William	The Well at the World's End	1896
morris_wood	Morris, William	The Wood Beyond the World	1894
Morrison_Investigator	Morrison, Arthur George	Martin Hewitt, Investigator	1894
Morrison_Triangle	Morrison, Arthur George	The Red Triangle	1903
Nabokov_Ada	Nabokov, Vladimir	Ada or Ardor: A Family Chronicle	1969
Nabokov_Harlequins	Nabokov, Vladimir	Look at the Harlequins!	1974
Nabokov_Knight	Nabokov, Vladimir	The Real Life of Sebastian Knight	1941
Nabokov_Lolita	Nabokov, Vladimir	Lolita	1955
Nabokov_Pnin	Nabokov, Vladimir	Pnin	1957
Nabokov_Sinister	Nabokov, Vladimir	Bend Sinister	1947
Nabokov_Transparent	Nabokov, Vladimir	Transparent Things	1972
Obrien_Policeman	O'Brien, Flann	The Third Policeman	1939-

			1940
Obrien_Swim	O'Brien, Flann	At Swim-Two-Birds	1939
Orwell_Aspidistra	Orwell, George	Keep the Aspidistra Flying	1936
Orwell_Burmese	Orwell, George	Burmese Days	1934
Orwell_Coming	Orwell, George	Coming Up for Air	1939
Orwell_Daughter	Orwell, George	A Clergyman's Daughter	1935
Orwell_Farm	Orwell, George	Animal Farm	1945
Orwell_Nineteen	Orwell, George	Nineteen Eighty-Four	1949
Peacock_Headlong	Peacock, Thomas Love	Headlong Hall	1915
Peacock_Marian	Peacock, Thomas Love	Maid Marian	1822
Peacock_Nightmare	Peacock, Thomas Love	Nightmare Abbey	1818
Poe_Letter	Poe, Edgar Allan	The Purloined Letter	1844
Porter_Chiefs	Porter, Jane	The Scottish Chiefs	1810
Porter_Thaddeus	Porter, Jane	Thaddeus of Warsaw	1803
Post_Sleut	Post, Melville Davisson	The Sleuth of St. James Square	1920
Post_Uncleabner	Post, Melville Davisson	Uncle Abner, Master of Mysteries	1918
Powys_Weymouth	Powys, John Cowper	Weymouth Sands	1934
Powys_Wolf	Powys, John Cowper	Wolf Solent	1929
Purbeck_Honoria	Purbeck, Elizabeth and Jane	Honoria Somerville: a novel	1789
Radcliffe_Sicilian	Radcliffe, Ann	A Sicilian Romance	1790
Radcliffe_Udolpho	Radcliffe, Ann	The Mysteries of Udolpho	1794
Drichardson_Interim	Richardson, Dorothy	Interim	1920
Drichardson_Pointe droofs	Richardson, Dorothy	Pointed Roofs	1915
Drichardson_Tunnel	Richardson, Dorothy	The Tunnel	1919
Rowling_Chamber	Rowling, J. K.	Harry Potter and the Chamber of Secrets	1998
Rowling_Goblet	Rowling, J. K.	Harry Potter and the Goblet of Fire	2000
Rowling_Hallows	Rowling, J. K.	Harry Potter and the Deathly Hallows	2007
Rowling_Order	Rowling, J. K.	Harry Potter and the Order of the Phoenix	2003
Rowling_Prince	Rowling, J. K.	Harry Potter and the Half-Blood Prince	2005
Rowling_Prisoner	Rowling, J. K.	Harry Potter and the Prisoner of Azkaban	1999

Rowling_Stone	Rowling, J. K.	Harry Potter and the Philosopher's Stone	1997
Rowling_Casual	Rowling, J. K.	The Casual Vacancy	2012
Salinger_Catcher	Salinger, Jerome David	The Catcher in the Rye	1951
Richardson_Clarissa	Samuel, Richardson	Clarissa, or, the History of a Young Lady	1748
Richardson_Grandison	Samuel, Richardson	The History of Sir Charles Grandison	1753
Richardson_Pamela	Samuel, Richardson	Pamela, or Virtue Rewarded	1740
Scott_Ivanhoe	Scott, Sir Walter	Ivanhoe	1819
Scott_Kenilworth	Scott, Sir Walter	Kenilworth	1821
Scott_Lammermoor	Scott, Sir Walter	The Bride of Lammermoor	1819
Scott_Rob	Scott, Sir Walter	Rob Roy	1817
Scott_Waverley	Scott, Sir Walter	Waverley	1814
Selden_Villasantelle	Selden, Catherine	Villa Santelle, or The Curious Impertinent	1817
Shelley_Frankenstein	Shelley, Mary	Frankenstein: or, The Modern Prometheus	1818
Shelley_Lastman	Shelley, Mary	The Last Man	1826
Shelley_Mathilda	Shelley, Mary	Mathilda	1819
Shelley_Valperga	Shelley, Mary	Valperga; or, The Life and Adventures of Castruccio, Prince of Lucca	1823
Sinclair_Jungle	Sinclair, Upton	The Jungle	1906
Sinclair_Samuel	Sinclair, Upton	Samuel The Seeker	1910
Smollett_Clinker	Smollett, Tobias George	The Expedition of Humphry Clinker	1771
Smollett_Pickle	Smollett, Tobias George	The Adventures of Peregrine Pickle	1751
Smollett_Random	Smollett, Tobias George	The Adventures of Roderick Random	1748
Spence_Curate	Spence, Elizabeth Isabella	The Curate and His Daughter: A Cornish Tale	1813
Stein_Lives	Stein, Gertrude	Three Lives	1905-1906
Stein_Toklas	Stein, Gertrude	The Autobiography of Alice B. Toklas	1933
Steinbeck_Eden	Steinbeck, John Ernst	East of Eden	1952
Steinbeck_Grapes	Steinbeck, John Ernst	The Grapes of Wrath	1939
Steinbeck_Mice	Steinbeck, John Ernst	Of Mice and Men	1937
Stephens_Crock	Stephens, James	The Crock of Gold	1912
Stephens_Mary	Stephens, James	Mary, Mary	1912

stevenson_arrow	Stevenson, Robert Louis	The Black Arrow	1883
stevenson_catriona	Stevenson, Robert Louis	Catriona	1893
Stevenson_Jekyll	Stevenson, Robert Louis	The Strange Case of Dr Jekyll and Mr Hyde	1886
Stevenson_Kidnapped	Stevenson, Robert Louis	Kidnapped	1886
stevenson_island	Stevenson, Robert Louis	Treasure Island	1882
Strutt_Drelincourt	Strutt, Elizabeth	Drelincourt and Rodalvi	1807
Stuart_Toledo	Stuart, Augusta Amelia	Cava of Toledo; or, the Gothic Princess	1812
Swift_Gulliver	Swift, Jonathan	Gulliver's Travels	1726
Swift_Tub	Swift, Jonathan	A Tale of a Tub	1704
Taylor_Rachel	Taylor, Jane	Rachel: a Tale	1817
Thackeray_Barry	Thackeray, William Makepeace	The Luck of Barry Lyndon	1844
Thackeray_Pendennis	Thackeray, William Makepeace	The History of Pendennis	1850
Thackeray_Vanity	Thackeray, William Makepeace	Vanity Fair	1848
Tolkien_Hobbit	Tolkien, John Ronald Reuel	The Hobbit or There and Back Again	1937
Tolkien_Lord1	Tolkien, John Ronald Reuel	The Fellowship of the Ring	1954
Tolkien_Lord2	Tolkien, John Ronald Reuel	The Two Towers	1954
Tolkien_Lord3	Tolkien, John Ronald Reuel	The Return of the King	1955
Tolkien_Silmarillion	Tolkien, John Ronald Reuel	The Silmarillion	1977
Tomlins_Victim	Tomlins, Elizabeth Sophia	The Victim of Fancy	1787
Trollope_Barchester	Trollope, Anthony	Barchester Towers	1857
Trollope_Forgive	Trollope, Anthony	Can You Forgive Her?	1865
Trollope_Phineas	Trollope, Anthony	Phineas Finn	1869
Trollope_Prime	Trollope, Anthony	The Prime Minister	1876
Trollope_Redux	Trollope, Anthony	Phineas Redux	1874
Trollope_Thorne	Trollope, Anthony	Doctor Thorne	1858
Trollope_Warden	Trollope, Anthony	The Warden	1855
Twain_Finn	Twain, Mark	Adventures of Huckleberry Finn	1885
Twain_Innocents	Twain, Mark	The Innocents Abroad, or The New Pilgrims' Progress	1869
Twain_Pauper	Twain, Mark	The Prince and the Pauper	1881
Twain_Sawyer	Twain, Mark	The Adventures of Tom	1876

		Sawyer	
Twain_Yankee	Twain, Mark	A Connecticut Yankee in King Arthur's Court	1889
Walpole_Otranto	Walpole, Horace	The Castle of Otranto	1764
Waugh_Brideshead	Waugh, Evelyn	Brideshead Revisited	1945
Waugh_Dust	Waugh, Evelyn	A Handful of Dust	1934
Waugh_Flags	Waugh, Evelyn	Put Out More Flags	1942
Waugh_Officers	Waugh, Evelyn	Officers and Gentlemen	1955
Waugh_Vile	Waugh, Evelyn	Vile Bodies	1930
Wells_Invisible	Wells, Herbert George	The Invisible Man	1897
Wells_MenMoon	Wells, Herbert George	The First Men in the Moon	1901
Wells_Moreau	Wells, Herbert George	The Island of Doctor Moreau	1896
Wells_Time	Wells, Herbert George	The Time Machine	1895
Wells_War	Wells, Herbert George	The War of the Worlds	1898
Wells_WarAir	Wells, Herbert George	The War in the Air	1908
West_Judge	West, Rebecca	The Judge	1922
West_Return	West, Rebecca	The Return of the Soldier	1918
Wharton_Age	Wharton, Edith	The Age of Innocence	1920
Wharton_Frome	Wharton, Edith	Ethan Frome	1911
Wharton_Mirth	Wharton, Edith	The House of Mirth	1905
Wilde_Dorian	Wilde, Oscar	The Picture of Dorian Gray	1891
Wilkinson_Child	Wilkinson, Sarah Scudgell	The Child of Mystery	1808
Mason_Road	Woodley, Alfred Edward Mason	The Broken Road	1907
Mason_Villarose	Woodley, Alfred Edward Mason	At the Villa Rose	1910
Woolf_Acts	Woolf, Virginia	Between the Acts	1941
Woolf_Dalloway	Woolf, Virginia	Mrs Dalloway	1925
Woolf_Guineas	Woolf, Virginia	Three Guineas	1938
Woolf_Jacobs	Woolf, Virginia	Jacob's Room	1922
Woolf_Lighthouse	Woolf, Virginia	To the Lighthouse	1927
Woolf_Monday	Woolf, Virginia	Monday or Tuesday	1941
Woolf_Night	Woolf, Virginia	Night and Day	1919
Woolf_Orlando	Woolf, Virginia	Orlando	1928
Woolf_Room	Woolf, Virginia	Between the Acts	1941
Woolf_Voyage	Woolf, Virginia	The Voyage Out	1915
Woolf_Waves	Woolf, Virginia	The Waves	1931
Woolf_Years	Woolf, Virginia	The Years	1937

The incorrect groups resulting from the network clustering (as described in **Chapter 3**) are delimited by horizontal bars; the shading indicates authorial connection between neighbouring groups:

Table 4

Compton_Dolores
Compton_Family, Compton_Men, ConFord_Nature
Ford_ARingForNancy
ConFord_Inheritors, Ford_TheBenefactor
Ford_AnEnglishGirl, Ford_Soldier
Ford_Ladies, Ford_MrApollo
Ford_Crowned, Ford_PrivySeal, Ford_Queen, Ford_TheHalfMoon
Galsworthy_Pharisees
Galsworthy_Chancery, Galsworthy_Let, Galsworthy_Man, Galsworthy_Monkey
Hawthorne_Fanshawe
Hawthorne_Blithedale, Hawthorne_MarbleFaun, Hawthorne_Scarlet, Hawthorne_SevenGables
Huxley_Mortal
Huxley_Brave, Huxley_Crome, Huxley_Gaza, Huxley_Point
Lessing_Planet, Lessing_Sirian
Lessing_Ben, Lessing_Child, Lessing_City, Lessing_Dream, Lessing_General, Lessing_MaraDann, Lessing_Orkney
ConFord_Romance, Conrad_Agent, Conrad_ArrowGld, Conrad_Chance, Conrad_Duel, Conrad_EndTethr, Conrad_Falk, Conrad_Freya, Conrad_Heart, Conrad_Lord, Conrad_Nostromo, Conrad_Rescue, Conrad_Rover, Conrad_ShaddowL, Conrad_SmileFortune, Conrad_Typhoon, Conrad_Victory, Conrad_Western
Conrad_Almayer, Conrad_NiggerN, Conrad_OutcastI
Joyce_Dubliners, Joyce_Portrait
Joyce_Finnegans, Joyce_Ulysses, Obrien_Swim
Obrien_Policeman, Stephens_Crock, Stephens_Mary
Doyle_MaracotDeep, Doyle_TheLostWorld
Doyle_Adventures, Doyle_Lastbow, Doyle_Study, Doyle_TheHound
Fielding_Amelia, Fielding_Joseph, Fielding_Tom
Fielding_Shamela, Richardson_Clarissa, Richardson_Grandison, Richardson_Pamela
Greene_BurntOut, Greene_Havana
Greene_Brighton, Greene_Confidential, Greene_Ministry, Greene_Power
Golding_Rites
Golding_Inheritors, Golding_Lord, Golding_Spire, Passos_Initiation, Passos_Soldiers
Melville_Bartleby
Melville_Mobydick, Melville_Omoo, Melville_Redburn, Melville_Typee, Twain_Innocents
Mitchell_Gonewind, Montgomery_Gables, Montgomery_Ingleside, Montgomery_Island, Montgomery_Quoted
Orwell_Aspidistra, Orwell_Burmese, Orwell_Coming, Orwell_Daughter,

Orwell_Nineteen
Orwell_Farm, Sinclair_Jungle, Stein_Toklas
Edgeworth_Absentee, Edgeworth_Assistant, Edgeworth_Belinda, Edgeworth_Ennui, Edgeworth_Forester, Edgeworth_Vivian
Aubin_Charlotta, Bradford_Plymouth, Defoe_Cruzoe, Defoe_Moll, Defoe_Roxana, Defoe_Singleton, Edgeworth_Rackrent
Dickens_Barnaby, Dickens_Boz, Dickens_Cities, Dickens_Curiosity, Dickens_Nicholas, Dickens_Oliver, Dickens_Pickwick
Dickens_Chimes, Dickens_Christmas, Dickens_Cricket
Dickens_Bleak, Dickens_David, Dickens_Dombey, Dickens_Dorrit, Dickens_Edwin, Dickens_Expectations, Dickens_Hard, Dickens_Mutual
James_Bostonians, James_Portrait, James_Princess, James_Reverbera, James_Tragic
James_American, James_Confidence, James_Europeans, James_Roderick, James_Washington, James_Watch
James_Ambassadors, James_Awkward, James_Golden, James_Maisie, James_Poynton, James_Sacred, James_Wings
ABrontë_Agnes, ABrontë_Tenant, Austen_Emma, Austen_Mansfield, Austen_Northanger, Austen_Persuasion, Austen_Pride, Austen_Sense
Morris_Child, Morris_House, Morris_JohnBall, Morris_Plain, Morris_Roots, Morris_Story, Morris_Sundering, Morris_Water, Morris_Well, Morris_Wood
Morris_Signs, Peacock_Headlong, Peacock_Nightmare, Poe_Letter, Sjohnson_Rasselas, Sjohnson_Scotland, Smollett_Clinker, Smollett_Pickle, Smollett_Random, Sterne_Sentimental, Sterne_Tristram, Swift_Gulliver, Swift_Tub, Woolf_Guineas
Beckett_Malone, Beckett_Molloy, Beckett_Unnamable, Beckford_Vathek, Bennet_Babylon, Bennet_Helen, Bennet_Imperial, Bentley_Trent, Bowen_Friends, Bowen_Hotel, Cary_Johnson, Catton_Luminaries
Faulkner_Absalom, Faulkner_Light, Faulkner_Moses
Maugham_Bondage, Maugham_Hero, Maugham_Magician, Maugham_Moon
Faulkner_Dying, Faulkner_Sound, Green_Loving, Green_Party, Hemingway_Across, Hemingway_Bell, Hemingway_Farewell, Hemingway_Fiesta, Hemingway_Menwithout, Hemingway_Oldman, Maugham_Liza, Salinger_Catcher, Stein_Lives, Steinbeck_Grapes, Steinbeck_Mice, Twain_Finn
Stevenson_Arrow, Stevenson_Catriona, Stevenson_Jekyll, Stevenson_Kidnapped, Stevenson_Treasure, Twain_Pauper, Twain_Sawyer, Twain_Yankee
Nabokov_Ada, Nabokov_Harlequins, Nabokov_Lolita, Nabokov_Pnin, Nabokov_Sinister, Nabokov_Transparent
Tolkien_Hobbit, Tolkien_Lord1, Tolkien_Lord2, Tolkien_Lord3
Nabokov_Knight, Tolkien_Silmarillion, Wells_Invisible, Wells_MenMoon, Wells_Moreau, Wells_Time, Wells_War, Wells_WarAir, West_Judge, West_Return, Wharton_Age, Wharton_Frome, Wharton_Mirth, Wilde_Dorian, Wlewis_Condemned, Wlewis_Tarr, Woolf_Acts, Woolf_Dalloway, Woolf_Jacobs, Woolf_Lighthouse, Woolf_Monday, Woolf_Night, Woolf_Orlando, Woolf_Room, Woolf_Voyage, Woolf_Waves, Woolf_Years
Capote_Voices, Ellison_Invisible
Harperlee_Mockingbird, Sinclair_Samuel, Steinbeck_Eden

Mlewis_Bravo, Peacock_Marian, Walpole_Otranto
Godwin_Imogen, Mlewis_Monk, Porter_Chiefs, Porter_Thaddeus, Radcliffe_Sicilian, Radcliffe_Udolpho, Selden_Villasantelle, Shelley_Frankenstein, Shelley_Lastman, Shelley_Mathilda, Shelley_Valperga **CHAWTON HOUSE**: Charlton_Parisian, Craik_Stella, Harvey_Anything, Harvey_Tynemouth, Helme_Magdalen, Hunter_Unexpected, Johnson_Francis, Mackenzie_Irish, Mackenzie_Monmouth, Purbeck_Honoria, Spence_Curate, Strutt_Drelincourt, Stuart_Toledo, Tomlins_Victim, Wilkinson_Child
Beckford_Azemia, Godwin_Caleb **CHAWTON HOUSE**:Hughes_Caroline
CHAWTON HOUSE: Jacson_Isabella, Jacson_Things, Taylor_Rachel
Green_Romance**, CHAWTON HOUSE**: Bennett_Agnes, Cooper_CarolineHerbert, Foster_Corinna, Foster_Substance, Hatton_Lovers, Martin_Enchantress
Burney_Camilla, Burney_Cecilia, Burney_Darblay, Burney_Evelina, Burney_Wanderer, Goldsmith_Vicar **CHAWTON HOUSE**: ALewis_Vicissitudes, Canning_Offspring, Carver_OldWoman, Mathews_Simple

A.4. Polish corpus 100

A small-scale corpus of 100 Polish novels published between 1850-1940 by 35 authors, 3 books each (but for J. Godlewska, L. Godlewska, and M. Samozwaniec), including 52 books written by female and 48 by male writers.

Table 5

File-id	Author	Title	Date[4]
balucki_murzyn	Bałucki, Michał	Biały murzyn	1875
balucki_przebudzeni	Bałucki, Michał	Przebudzeni	1864
balucki_burmistrz	Bałucki, Michał	Pan burmistrz z Pipidówki (Powieść życia autonomicznego Galicji)	1887
berent_diogenes	Berent, Wacław	Diogenes w kontuszu	1937
berent_kamienie	Berent, Wacław	Żywe kamienie	1918
berent_prochno	Berent, Wacław	Próchno	1903
dabrowska_ nocednie1	Dąbrowska, Maria	Noce i dnie. T.1	1931
dabrowska_ nocednie2	Dąbrowska, Maria	Noce i dnie. T.2	1932
dabrowska_ nocednie3	Dąbrowska, Maria	Noce i dnie. T.3	1933
deotyma_panienka	Deotyma	Panienka z okienka	1893
deotyma_rozdrozu	Deotyma	na rozdrożu	1877
deotyma_zagadka	Deotyma	Zwierciadlana zagadka	1879
dmochowska_dwor	Dmochowska, Emma	Dwór w Haliniszkach	1903
dmochowska_ obraczka	Dmochowska, Emma	Obrączka	1907
dmochowska_ odlamana	Dmochowska, Emma	Jak odłamana	1914
mostowicz_hanki	Dołęga-Mostowicz, Tadeusz	Pamiętnik Pani Hanki	1939
mostowicz_kariera	Dołęga-Mostowicz, Tadeusz	Kariera Nikodema Dyzmy	1932
domanska_historia	Domańska, Antonina	Historia żółtej ciżemki	1913
domanska_krysia	Domańska, Antonina	Krysia bezimienna	1914
domanska_paziowie	Domańska, Antonina	Paziowie króla Zygmunta	1910
dygasinski_as	Dygasiński, Adolf	As	1896
dygasinski_ piszczalski	Dygasiński, Adolf	Pan Jędrzej Piszczalski, Opowieść z niedawnej	1890

[4] Date of publishing or, if known, of writing.

		przeszłości	
dygasinski_wilk	Dygasiński, Adolf	Wilk, psy i ludzie	1883
godlewska_ninka	Godlewska, Janina	Ninka	1897
godlewska_kato	Godlewska, Ludwika	KATO Powieść Współczesna	1897
godlewska_kwiat	Godlewska, Ludwika	Kwiat aloesu	1897
gojawiczynska_ dziewczeta	Gojawiczyńska, Pola	Dziewczęta z Nowolipek	1935
gojawiczynska_ ziemia	Gojawiczyńska, Pola	Ziemia Elżbiety	1934
gojawiczynska_ jablon	Gojawiczyńska, Pola	Rajska jabłoń	1937
beczkowska_droga	Grot-Bęczkowska, Wanda	Kędy droga?	1898
beczkowska_ gniezdzie	Grot-Bęczkowska, Wanda	W mieszczańskim gnieździe	1899
beczkowska_bedzie	Grot-Bęczkowska, Wanda	Co będzie z naszego chłopca?	1897
iwaszkiewicz_ czerwone	Iwaszkiewicz, Jarosław	Czerwone Tarcze	1934
iwaszkiewicz_mlyn	Iwaszkiewicz, Jarosław	Młyn nad Utratą	1936
iwaszkiewicz_panny	Iwaszkiewicz, Jarosław	Panny z Wilka	1932
kaczkowski_grob	Kaczkowski, Zygmunt	Grób Nieczui	1857
kaczkowski_ murdelio	Kaczkowski, Zygmunt	Murdelio	1853
kaczkowski_ olbrachtowi	Kaczkowski, Zygmunt	Olbrachtowi rycerze	1889
korzeniowski_ emeryt	Korzeniowski, Józef	Emeryt	1851
korzeniowski_ garbaty	Korzeniowski, Józef	Garbaty	1853
korzeniowski_ krewni	Korzeniowski, Józef	Krewni	1856
kossak_bog	Kossak, Zofia	Krzyżowcy	1935
kossak_oreza	Kossak, Zofia	Bez oręża	1937
kossak_zmilosci	Kossak, Zofia	Z miłości	1925
kraszewski_ kordecki	Kraszewski, Józef Ignacy	Kordecki	1850
kraszewski_lalki	Kraszewski, Józef Ignacy	Lalki, Sceny przedślubne	1874
kraszewski_piast	Kraszewski, Józef Ignacy	Król Piast (Michał Książę Wiśniowiecki), Powieść historyczna	1888

krzemieniecka_ fatum	Krzemieniecka, Hanna	Fatum	1904
krzemieniecka_ odejdzie	Krzemieniecka, Hanna	A gdy odejdzie w przepaść wieczną, Romans zagrobowy	1910
krzemieniecka_ wichry	Krzemieniecka, Hanna	Lecą wichry! : powieść	1923
kuncewiczowa_ cudzoziemka	Kuncewiczowa, Maria	Cudzoziemka	1936
kuncewiczowa_ ksiezyce	Kuncewiczowa, Maria	Dwa księżyce	1933
kuncewiczowa_ twarz	Kuncewiczowa, Maria	Twarz mężczyzny	1928
makuszynski_ basie	Makuszyński, Kornel	Awantura o Basię	1937
makuszynski_ drodze	Makuszyński, Kornel	Po Mlecznej Drodze	1917
makuszynski_ szalenstwa	Makuszyński, Kornel	Szaleństwa panny EWY	1940
marrene_bozek	Marrené, Waleria	Bożek Miljon	1871
marrene_mezowie	Marrené, Waleria	Mężowie i żony	1875
marrene_roza	Marrené, Waleria	Róża	1872
mniszek_gehenna	Mniszkówna, Helena	Gehenna czyli dzieje nieszczęśliwej miłości	1914
mniszek_ordynat	Mniszkówna, Helena	Ordynat Michorowski	1910
mniszek_tredowata	Mniszkówna, Helena	Trędowata	1909
mostowicz_murek	Mostowicz, Dołęga, Tadeusz	Doktor Murek	1936
nalkowska_granica	Nałkowska, Zofia	Granica	1935
nalkowska_kobiety	Nałkowska, Zofia	Kobiety	1906
nalkowska_romans	Nałkowska, Zofia	Romans Teresy Hennert	1923
orzeszkowa_gloria	Orzeszkowa, Eliza	Gloria victis	1910
orzeszkowa_meir	Orzeszkowa, Eliza	Meir Ezofowicz	1878
orzeszkowa_ niemnem	Orzeszkowa, Eliza	Nad Niemnem	1888
prus_emancypantki	Prus, Bolesław	Emancypantki	1894
prus_faraon	Prus, Bolesław	Faraon	1897
prus_lalka	Prus, Bolesław	Lalka	1890
reymont_chlopi	Reymont, Władysław Stanisław	Chłopi	1908
reymont_ komediantka	Reymont, Władysław Stanisław	Komediantka	1896
reymont_obiecana	Reymont,	Ziemia obiecana	1899

	Władysław Stanisław		
rodziewicz_lato	Rodziewiczówna, Maria	Lato leśnych ludzi	1920
rodziewicz_miedzy	Rodziewiczówna, Maria	Między ustami a brzegiem pucharu	1890
rodziewicz_straszny	Rodziewiczówna, Maria	Straszny dziadunio	1887
samozwaniec_ ustach	Samozwaniec, Magdalena	Na ustach grzechu	1922
sienkiewicz_ogniem	Sienkiewicz, Henryk	Ogniem i mieczem	1884
sienkiewicz_quo	Sienkiewicz, Henryk	Quo vadis	1896
sienkiewicz_rodzina	Sienkiewicz, Henryk	Rodzina Połanieckich	1894
sygietynski_ calvados	Sygietyński, Antoni	Na skałach Calvados	1884
sygietynski_ogien	Sygietyński, Antoni	Święty ogień	1918
sygietynski_ wysadzony	Sygietyński, Antoni	Wysadzony z siodła	1891
swietochowski_ drygalowie	Świętochowski, Aleksander	Drygałowie	1914
swietochowski_ prawdy	Świętochowski, Aleksander	Tragikomedya prawdy	1888
swietochowski_ twinko	Świętochowski, Aleksander	Twinko	1936
zapolska_smierc	Zapolska, Gabriela	Śmierć Felicyana Dulskiego	1911
zapolska_kaska	Zapolska, Gabriela	Kaśka Kariatyda	1888
zapolska_tagiejew	Zapolska, Gabriela	Pan policmajster Tagiejew	1905
zarzycka_dzikuska	Zarzycka, Irena	Dzikuska	1927
zarzycka_irka	Zarzycka, Irena	Panna Irka	1931
zarzycka_wiatr	Zarzycka, Irena	Pod wiatr	1934
zeromski_bezdomni	Żeromski, Stefan	Ludzie bezdomni	1899
zeromski_ przedwiosnie	Żeromski, Stefan	Przedwiośnie	1924
zeromski_syzyfowe	Żeromski, Stefan	Syzyfowe prace	1897
zulawski_laus	Żuławski, Jerzy	Laus feminae	1914
zulawski_srebrnym	Żuławski, Jerzy	Na srebrnym globie	1903
zulawski_zwyciezca	Żuławski, Jerzy	Zwycięzca	1910

A.5. EN-PL parallel corpus

The sample corpus comprises 39 books of one English author only, Terry Pratchett. The corpus of translations into Polish is short by one (crossed out, in the table below), and is authored by one translator, Piotr Cholewa (other translations, by Dorota Malinowska-Grupińska, are neglected). The bibliographic information has been copied from Wikipedia (2014).

Table 6

Original title	Date	Polish title	Date
The Colour of Magic	1983	Kolor magii	1994
The Light Fantastic	1986	Blask fantastyczny	1995
Equal Rites	1987	Równoumagicznienie	1996
Mort	1987	Mort	1996
Sourcery	1988	Czarodzicielstwo	1997
Wyrd Sisters	1988	Trzy wiedźmy	1998
Pyramids	1989	Piramidy	1998
Guards! Guards!	1989	Straż! Straż!	1999
Eric	1990	Eryk	1997
Moving Pictures	1990	Ruchome obrazki	2000
Reaper Man	1991	Kosiarz	2001
Witches Abroad	1991	Wyprawa czarownic	2001
Small Gods	1992	Pomniejsze bóstwa	2001
Lords and Ladies	1992	Panowie i damy	2002
Men at Arms	1993	Zbrojni	2002
Soul Music	1994	Muzyka duszy	2002
Interesting Times	1994	Ciekawe czasy	2003
Maskerade	1995	Maskarada	2003
Feet of Clay	1996	Na glinianych nogach	2004
Hogfather	1996	Wiedźmikołaj	2004
Jingo	1997	Bogowie, honor, Ankh-Morpork	2005
The Last Continent	1998	Ostatni kontynent	2006
Carpe Jugulum	1998	~~Carpe Jugulum~~	~~2006~~
The Fifth Elephant	1999	Piąty elefant	2006
The Truth	2000	PRAWDA	2007
Thief of Time	2001	Złodziej czasu	2007
The Last Hero	2001	Ostatni bohater	2003
The Amazing Maurice and his Educated Rodents	2001	Zadziwiający Maurycy i jego edukowane gryzonie	2004
Night Watch	2002	Straż nocna	2008
The Wee Free Men	2003	Wolni Ciut Ludzie	2005
Monstrous Regiment	2003	Potworny regiment	2008
A Hat Full of Sky	2004	Kapelusz pełen nieba	2005
Going Postal	2004	Piekło pocztowe	2008
Thud!	2005	Łups!	2009
Wintersmith	2006	Zimistrz	2007
Making Money	2007	Świat finansjery	2009
Unseen Academicals	2009	Niewidoczni Akademicy	2010
I Shall Wear Midnight	2010	W północ się odzieję	2011
Snuff	2011	Niuch	2012

Appendix B: Software and parameters

In this appendix I provide a few hints as to what software I used and what were the parameters used for the particular studies, which can be helpful for readers willing to conduct similar research. All the necessary and sufficient information to run the programs is contained in the respective read-me files and how-tos.

Some general information on the workings of PoS taggers can be found in the subsection **Morpho-syntactic annotation** of Ide (2004).

B.1. Tag-sets

The common English-Polish tag-set I used is based on the Penn Treebank tag-set (Santorini 1990), in comparison to which some of the categories are additionally merged or omitted; being aware of the "unfortunate fact that it is often extremely difficult and sometimes impossible to map one tag-set to another, which has resulted in much re-creation of lexical information to suit the needs of a particular tagger" (Ide 2004), one has to say that in creating many-language tag-sets such drawbacks are inevitable. The modifications are to a much extent arbitrary and should be treated only as a working version.

The tables below present replacement rules from Penn Treebank tag-set and from the Polish tag-set used in TaKIPI (Woliński 2003) to the common EN-PL tag-set ("-" means omission of a category; asterisk, e.g., in "subst:sg:*", means including all other subcategories, in this case all nouns in the singular irrespective of their case, declension type, etc.). While I am not so much interested in comparing the grammatical structures of the two languages, but rather in how the translation between them is performed, some of the replacement rules do not reflect "equivalence" between parts of speech, but rather are motivated by the intuition on what happens with some structures in translation (e.g., with existential *there* or with the determiners in EN->PL translation). Some of the most disputable solutions are discussed in the footnotes. The rest is, hopefully, fairly self-explanatory.

Penn Treebank	Common EN-PL
#	SYM
$	SYM
-LSB-	-LRB-
-RSB-	-RRB-
DT	-[5]
EX	PRP
LS	-
MD	VBP
NNPS	NNP
NNS	NN
PDT	JJ
POS	-[6]
PRP$	PRP
RP	-[7]
TO	IN
VBZ	VBP
WDT	CC
WP	CC
WP$	CC
WRB	CC

TaKIPI	Common EN-PL
adj*comp	JJR
adj*pos	JJ
adj*sup	JJS
adja	JJ
adv:comp	RBR
adv:pos	RB
adv:sup	RBS
aglt*	PRP[8]
bedzie*	VBP
conj	CC[9]
depr:pl:*	NNP
depr:sg:*	NN
fin*	VBP
ger*	VBG
imps*	VBN
impt*	VB
inf*	VB
interp	.[10] , " : -LRB- -RRB-
num*	CD[11]
pact*	VBG
pant*	VBN
pcon*	VBG
ppas*	VBN

[5] The determiners (DT) are omitted, since they are usually non-existent in Polish their rough equivalents: “jakiś”, “taki”, etc., belong to the group adj:...:pos which is replaced by JJ. For consistency, predeterminers (PDT) are also replaced with JJ.

[6] The possessive ending (POS) is omitted, since in Polish the possessive would be grammaticalised as genitive case and so it does not obtain a separate category tag.

[7] RP is a particle, i.e., roughly the preposition in a phrasal verb in English, whose function in Polish is fulfilled by a prefix of a verb. The omission allows to just count verbs without delving into their morphological constituents.

[8] This is a fairly unintuitive replacement: the agglutinative “to be” in Polish, e.g., in “wlazł[praet:sg:m1:perf]eś[aglt:sg:sec:imperf:wok]” translated as “[you have] climbed” could be omitted, since it is morphologically integrated with the verb, but I decided to replace it with a personal pronoun (PRP) category, as it carries the grammatical information on person, which in English would be rendered as an obligatory pronominal subject of a verb.

[9] This is the major problem with adjusting the Penn Treebank tag-set: CC contains only the coordinating conjunctions, while the subordinating ones belong to IN, and additional processing for their extraction would be needed. This affects the comparative benchmark distributions of PoS tags between EN and PL, as can be seen in Fig. …

[10] Punctuation in TaKIPI is all flattened to just one tag, but since it carries some statistical information on clause structures, I expanded it back to the Penn Treebank categories.

[11] The problem with this replacement is that Penn Treebank tags ordinal numbers with JJ, and additional processing for their extraction would be needed.

ppron12:*	PRP
ppron3:*	PRP
praet*	VBD
pred*	VBP
prep*	IN[12]
prep+adjp	RB[13]
qub	UH[14]
siebie*	PRP
subst:pl:*	NNP
subst:sg:*	NN
tmail	-
tnum*	CD
tsym	SYM
ttime	CD
turi	-
winien*	VBP
xxs	FW
xxx	FW

B.2. English PoS taggers

In order to perform part-of-speech tagging in English I used *Stanford Log-linear Part-Of-Speech Tagger* (Toutanova and Manning 2000, Toutanova et al. 2003) implemented in Java. I did not train the tagger myself, but used the *english-left3words-distsim.tagger* model for English, which was trained on PennTreebank's *Wall Street Journal* sections 0-18 and extra parser training data using the *left3words* architecture and includes word shape and distributional

similarity features. It is the default model of the Stanford NLP group because of its speed. The trained tagger can be downloaded from: http://nlp.stanford.edu/software/tagger.shtml.

The *Stanford Tagger* uses PennTreebank tag-set (Santorini 1990). The accuracy of the model on the standard *Wall Street Journal* 22-24 test set is around 97% (around 90% on unknown words).

[12] The problem is connected to the one in Footnote 9: IN incorporates both prepositions and subordinating conjunctions.

[13] Although it is only a partial solution, whenever a postprepositional adjective is found (see discussion in (Woliński 2003), it is joined with the preceding preposition, e.g., "po polsku" or "po prostu", and jointly rendered as an adverb (RB), to which it seems to be functionally equivalent. The replacement is disputable, however, as in translation it may be rendered as a prepositional phrase.

[14] The lexical range of the two tags is unfortunately quite different, as can be expected from Fig. … . The simplest interjections (UH) supposedly could be tagged as qublics (roughly, adverbial particles, which are not subject to any inflections), but the qublics contain also such words as *czy, nigdy, już, się, tam, nie,* etc., some of which could be tagged in English as adverbs, determiners, pronouns or finally interjections, depending on the precise use.

B.3. Polish PoS tagger

In order to perform part-of-speech tagging in Polish I used *TaKIPI 1.8* Polish language tagger (Piasecki 2007), whose name comes from Korpus Instytutu Podstaw Informatyki Polskiej Akademii Nauk (Korpus IPI PAN, or even shorter KIPI), i.e., the Corpus of the Institute of Computer Science of the Polish Academy of Sciences. It can be downloaded from the following website: http://nlp.pwr.wroc.pl/takipi/.

TaKIPI works based on the tag-set of the KIPI corpus (Woliński 2003). To be operational it needs the contextless morphological analyser *Morfeusz* [Morfeusz], which segments the tokens found in a text into their elementary morphological components, and thus allows to find a range of possible lexemes to which the token belongs. It can be downloaded from: http://sgjp.pl/morfeusz/. Beside the morphological analysis of known tokens done by *Morfeusz, TaKIPI* performs morphological disambiguation, and also contains a subroutine that guesses the morphosyntactic structure of the tokens unknown to it. Its authors claim 93.4% accuracy in tagging all tokens in a text.

I used the version of *Morfeusz* based on the *Polimorf* dictionary; the choice of the version affects the range of tokens covered. As concerns disambiguations, for the purpose of this thesis I always took only the first, i.e., the most probable form.

B.4. Stylo *R package*

The *Stylo* package (Eder et al. 2013) can be downloaded from: https://sites.google.com/site/computationalstylistics/scripts.

The accuracy of authorship attribution on an EN100 Benchmark corpus is around 96%-98%, depending on the classifier/clustering method, as presented in **Figure 14**.

The results presented in **Chapter 3**, were all performed with the use of classic Burrows's delta, either with the option *Cluster Analysis* (generating standard dendrograms) or *Consensus Tree*. The precise parameters are given below each figure. Pronouns were deleted in order to control for the narration type. So called *culling* parameter specifies the minimum percentage of the corpus in which the features used should appear (in the trivial case of 0% there is no limitation, while in the case of 100% the features should appear at least once in all the texts in the corpus). All these options can be easily chosen using the built-in Graphical User Interface. This part was done still using version 0.4.9.2 of the package.

The results presented in **Figure 14**, EN100 Benchmark, were obtained using version 0.5.6, without the use of GUI. The primary (training) set contained 2 texts per author and the secondary (test) set 1 text per author. The function classify() took option cv.folds=100 for cross-validation. For EN500 corpus no cross-validation was performed, since I had at my disposal only the distance table, not the frequency tables nor the texts themselves.

B.5. Community detection algorithms

Some preliminary clustering of the small English benchmark corpus was performed with *Infomap* software package for multi-level network clustering (Edler and Rosvall 2013), with the options –u (undirected network, since the Delta distances are symmetric) and –N 100 (i.e., the number of trials to run before picking the best solution), but resulted in just one big group comprising all the books.

The method I used further was the Louvain method of modularity maximisation (Blondel et al. 2008). The code was developed by Traag et al. (2011), which can be downloaded from:
https://launchpad.net/louvain.

To produce the **Figure 13 and Figure 15** the Delta distances were transformed into similarities as

$$s_{ba} = \delta_{ba}{}^{-p},$$

with the power exponent *p* ranging from 3 to 6 in increments of 0.25; the range of MFW checked was 100-4400 in increments of 100, and culling 0-100% in increments of 20%. The maximum Normalised Mutual Information scores were obtained for the power 5.5, 300 MFW at 100% culling, and the modularity resolution of 3.5. These parameters were used to produce the Figures mentioned above. The functional relation of distance and similarity has been chosen from among several different ones, so that it optimized the results of clustering was. An alternative could be a logarithmic relation (Eder 2014).

In order to produce **Figure 16** the Delta distances were transformed with the same power exponent $p = 5.5$ (so that EN100 could be treated as "training" for EN500 corpus), and 190 MFW were taken at 100% culling; the modularity resolution parameter scanned ranged from 3.0 to 9.0 in increments of 0.1, which resulted in the different numbers of authorial groups as seen in the Figure.

Abstract

The topic of this thesis is the computational methods for measurement of authorial style and algorithms of authorial attribution.

The first aim of the thesis was an attempt at a quantifiable separation of various layers of authorial style (in the present case the lexical and grammatical layers) in order to estimate their influence on the results of a chosen method of authorial attribution. Within the scope of these studies I compared the distance, so called Burrows's Delta, between a pair of English novels by two chosen authors and automatically generated texts, whose statistical distributions of parts of speech were borrowed from one of the authors, while the vocabulary from the other one; additionally, in the computatrificial texts I left the sets of words of the first author if they belonged to a particular part of speech. Such procedure allowed to create a hybrid text, which was attributed to the first author, even though the majority of lexical items were that of the second author.

The second aim was to identify the influences of the style and language of the original on the style of the translation. This part of research involved among others adapting Polish and English part-of-speech tag-sets to form a common translatorial tag-set. Beside making a couple of simple observations concerning the distributions and coocurrences of parts of speech in the two languages, I managed to determine some features of the selected translatorial corpus, which lie on the fringes of what seems a norm for Polish.

The third aim was testing the accuracy of state-of-the-art (unsupervised) clustering methods for automatic grouping of texts according to their author. The results show that the methods recognise authorship worse than the known supervised machine learning methods.

In the thesis I made use of corpora totalling around 550 digitised English-language novels and 100 Polish ones, as well as a parallel corpus of 39 novels of a single English author together with their translations by a single Polish translator. The research conducted involved utilising existing part-of-speech taggers (both for English and Polish), authorship attribution programmes, and programmes for graph clustering.

Streszczenie

Przedmiotem niniejszej pracy są komputerowe metody pomiaru stylu autorskiego oraz algorytmy rozpoznawania autorstwa tekstu.

Pierwszym z jej celów była próba kwantyfikowalnego rozdzielenia różnych warstw stylu autorskiego (w tym wypadku: warstwy leksykalnej i gramatycznej), aby oszacować ich wpływ na wyniki wybranej metody atrybucji autorskiej. W ramach tych badań porównywałem odległość, tzw. Deltę Burrowsa, pomiędzy rzeczywistymi powieściami dwu wybranych autorów oraz tekstami wytwarzanymi automatycznie, mającymi statystyczne rozkłady części mowy zapożyczone od jednego z autorów, słownictwo zaś od drugiego; dodatkowo w tekstach sztucznych pozostawiałem zestawy słów odpowiadających konkretnym częściom mowy zebrane z tekstu pierwszego autora. Pozwoliło to na stworzenie tekstu-hybrydy, któremu pomimo zachowania przeważającej części warstwy leksykalnej drugiego z autorów, przypisywane było autorstwo autora pierwszego.

Drugim celem było wstępne rozpoznanie wpływu stylu i języka oryginału na styl tłumaczenia. Ta część pracy wymagała m.in. roboczego opracowania wspólnego, polsko-angielskiego zestawu znaczników części mowy. Poza poczynieniem prostych obserwacji dot. rozkładów oraz współwystępowania części mowy w obu językach, udało się znaleźć pewne cechy wybranego zbioru przekładów, które lokują się na obrzeżach normy dla polszczyzny.

Trzecim celem było sprawdzenie skuteczności nowoczesnych (nienadzorowanych) metod analizy skupień w automatycznym grupowaniu tekstów wedle ich autorstwa. Wyniki wskazują, że metody te rozpoznają autorstwo gorzej niż metody nadzorowane znane dotychczas.

W pracy wykorzystane zostały korpusy o łącznej liczbie ok. 550 zdigitalizowanych powieści anglojęzycznych, oraz 100 powieści polskich, a także korpus równoległy 39 powieści jednego autora anglojęzycznego wraz z ich tłumaczeniami na język polski wykonanymi przez jednego tłumacza. Przeprowadzone badania wymagały użycia już istniejących programów do oznaczania części mowy (zarówno w polszczyźnie, jak i angielszczyźnie), programów do atrybucji autorskiej oraz programów do analizy skupień na grafach.