Fotis Jannidis / Gerhard Lauer

Burrows Delta and its Use in German Literary History

Abstract

We explore the use of Burrows's delta for research in German literary history. First we outline recent trends in quantitative and corpus based stylometry as a way of distant reading, before we discuss more closely the methodology of Burrows's delta and his further development by Eder's and Rybicki's stylometry with R. In a series of tests we analyze the German literature around 1800 a) according their authorship, epoch, genre, and gender. We look more closely on b) one individual style, that of Heinrich von Kleist. Our purpose is not to give new insights in German literary history but to validate a stylometric approach in literary history. As we can make plausible stylometry is a useful quantitative method for research on literary history. Some follow-ups are discussed, which go beyond common ways of doing literary history.

Indexing terms

Burrows's delta, corpus stylistics, literary history, statistic of literature, quantitative approach, stylometry with R, word frequency

Charlotte v. Ahlefeld, Ferdinand Buchholz, John F. Burrows, Gregory Crane, Caroline de la Motte Fouqué, Maciej Eder, Joseph v. Eichendorff, Stanley Fish, Johann Wolfgang Goethe, Hermann Hettner, Friedrich Hölderlin, David L. Hoover, Therese Huber, August Wilhelm Iffland, Jean Paul, Heinrich von Kleist, August Kotzebue, August Lafontaine, Thomas Corwin Mendenhall, Franco Moretti, August de Morgan, Martin Mueller, Benedikte Naubert, Novalis, Jan Rybicki, Friedrich Schiller, August Wilhelm Schlegel, Dorothea Schlegel, Friedrich Schlegel, Ludwig Tieck, Friederike Helene Unger, David Wellbery, Johann Karl Wezel

1. Introduction

Literary history is a persuading guide to the canon of the great books. It doesn’t matter whether you take a history of literature from nineteenth century in hand or a modern one, they all tend more or less to expose a small list of books out of the ocean of the published. For the year 1809 any history of German literature highlights Goethe’s novel *Wahlverwandtschaften [Elective Affinities]* as no other literary book seems to have be printed. This is the case regardless whether one consults the literary history of Hermann Hettner from 1870 or David Wellbery's from 2005.[1] But a short view on book catalogues around 1809 revises this canonical picture of literary history. Around hundred German novels were published in 1809, to mention widely read novels e.g. by August Lafontaine, his third volume of *Die beiden Bräute [The two brides]*, or August Kotzebue’s *Philbert oder die Verhältnisse [Philbert or the circumstances]*. Canon is one story, another story is the cultural history of read books. If literary history is more than the history of the great books a corpus based approach could be one way to deal with the high number of circulating books. ‘High number’ is not a metaphor especially if we talk about the literary history of 19th century and even there just for the German speaking countries. A serial like *Das belletristische Ausland [The belletristic foreign countries]*, published between 1843–1865, contains 3.618 volumes of translated books for the German literary reading audience.[2] Only in the year 1871 nearly a thousand new belles lettres books were printed. And in 1913 more than five thousand prose and poetry books came out.[3] The corpus of 19th century German books is huge and--as Gregory Crane wrote in his seminal paper “What do you do with a million books?”[4]--life is too short to read only what was published in a single year of the long nineteenth century.

The number of texts and the ability to access them has always had an influence on philological studies and new access to sources has always been a game changing event for philologists. When the humanists of the 15th and 16th century searched the libraries of monasteries for forgotten classical manuscripts, their findings were the basics of a new understanding of history, natural science and the humanities. When the founding fathers of the new university-based philology, like the brothers Grimm, did their research around 1800, access to the newly evaluated sources of medieval times was highly cherished and blocked for possible rivals. And today we have a similar situation with the new access to larger and large collections of literary texts. In the context of what we call nowadays Digital Humanities quantitative studies of literary texts have been repeatedly done, but only in the last approximately 5 years the amount of texts available has allowed large scale research in a qualitative new way. For German texts the time is even shorter because just now with the publishing of the *Digitale Bibliothek* by TextGrid the large collection of texts edited by *DirectMedia* and later posted online via Zeno.org has been provided with metadata which makes them useable for corpus research.[5]

Together with the growing power of the computer the exploding number of digital available texts alters the landscape of humanities scholarship in more than one respect. Latest when Google digitized about 20 million books from the estimated total number of 130 million unique books in the world, published in 480 languages,[6] books become a corpus. The existence of large corpora gave rise to a new interest in tools to extract information from them and with Google's ngram viewer,[7] a popular example of what kind of research may be possible, was widely discussed and though the obvious weaknesses of this approach were quickly pointed out,[8] it will stay to be one of the first examples of a tool which allows every humanist and not just the specialist to put his or her question to a vast corpus of texts. It is part of this ongoing change that common reading practices are under discussion. It is obvious that millions of books can’t be read in known methodologies of close reading. With the term “distant reading” Franco Moretti coined in 2000 a different approach and claimed that “literary history will quickly become very different from what it is now: it will become ‘second hand’: a patchwork of other people’s research, *without a single direct textual reading*. Still ambitious, and actually even more so than before (world literature!); but the ambition is now directly proportional *to the distance from the text*: the more ambitious the project, the greater must the distance be“.[9] These are for sure polemic words to mark a shift in scholarship where distant is a chance but for the price of losing a familiarity with books. No wonder that critic by Stanley Fish[10] and others attack Moretti for given up the intimacy with books as friends and to trade interpretation in for nothing more than patterns.

The debate on distant reading is more a moral one than a discussion on methodology. As Martin Mueller has shown by his term of “scalable reading”[11] there is no such thing as an opposition between close and distant reading. Only varied ways of looking on different amount of texts discriminate the one from the other view. It depends on your research questions at which point of the close to distance scale we read. History of literature is obviously this kind of research with more distance to single texts than the interpretation of Goethe’s *Wahlverwandtschaften [Elective Affinities]*. And not only this part of the debate around distant reading is misleading. Unlike the debates around Moretti’s term suggest there is a long scholarly tradition in literary studies and linguistics, which makes use of techniques like counting words or calculating patterns. Tradition of lexicometry, research on authorship attribution, stemmatology, concordance, or phylometry are established, although small areas of humanities scholarship since more than hundred years. The computer now offers to go some step further but in the shoes of former scholarship. A methodology behind Google ngram viewer explores cultural trends on a corpus of not less than 4% of the books ever printed.[12] This is an amount of books no one before the age of digitalization could handle. And Google's ngram viewer is not the only new tool for a quantitative and corpus based research. Machine learning by having trained on a set of learning examples to work on unseen data[13] or topic modeling,[14] which identifies shared themes through statistical models, social network analysis[15] for actor pattern recognition and others more are today's methodology to answer old questions in the humanities.

Stylometry is one of them, and as other more quantitative approaches it also has also a longer history. Stylometry is more or less part of research on authorship attribution and was used mostly on small corpora of books. In 1851 it was the mathematician August de Morgan who first proposed the statistical average of word length, measured in syllable, as criteria to discriminate authorships.[16] His test case was the epistles of Saint Paul, but soon others like Thomas Corwin Mendenhall followed him with his word length studies on Shakespeare/Marlowe. After 1960 linguists expanded the methodology on larger corpora and made increasing use of new characteristics like average sentence length, the vocabulary richness by type:token ratio, the homogeneity of texts by the number of new word types within a text, syntactic features and word classes.[17] Author attribution has become more and more a branch of corpus linguistics by using a growing range of measures and more and more complex correlation of these factors.[18] Today stylometry makes use of multivariant methods, these are primarily cluster analysis, correspondence analysis and principle component analysis. By cluster analysis texts are grouped according to their similarity with regard to features like word classes or word frequencies or other given features. By correspondence analysis more than one text feature is analyzed in a two dimensional matrix. And by principle component analysis correlations between a much larger range of variables are group in sets—so called ‘component’ or ‘dimension’--that show the most correspondences. All these kinds of analysis rely on lexical rather that semantic discriminators and as a rule of thumb they all work better if the corpus of texts is large. Tools like ‘Voyant’[19] by Stefan Sinclair and Geoffrey Rockwell offer these new kinds of text analysis, which starts to alter the way how philology is working in the next years.

2. Methods

Since his groundbreaking paper from 2002[20] John F. Burrows’s delta is a common standard in stylometry to measure the relative stylistic difference between two or more texts. Burrows developed his method in the context of author attribution studies in order to allow a selection of a few good candidates out of a larger group of possible candidates for the authorship of an anonymous text. In the beginning it was basically meant to reduce a large set to a small group in order to allow more complex and time consuming procedures to be applied to this smaller group. The procedure is fairly simple. As much of Burrows other rightly famous work it is based on counting the most common words--without the use of a stopword list, so especially the top of the list consists of words which have almost no semantics at all, like ‘the’, ‘is’, ‘a’ etc. Burrows based the practical demonstration of Delta on longer epic poems, so it was of interest to see in the following years studies which repeated the high success rate of correct attributions with other sorts of texts, mainly novels,[21] and also in other languages.[22] Today we can say that at least for some languages like English Burrows Delta has proven to be a remarkably good indicator of stylistic affinity. Not to get mislead: It is nothing like a ‘stylistic fingerprint' or 'stylistic DNA' or whatever other forensic comparison is used, because Burrows's delta doesn't identify an individual but describes the relation between a text and other texts in the context of this group of texts. So if you change the group you will also see new relations and maybe new clusterings.

However as Burrows himself and follow up studies by David L. Hoover and others have shown, Delta works well with texts longer than 2000 words with high probabilities to indicate the likely author – if there are other texts of the author part of the comparison. Some differences in the validity of results could be found working on prose and on poetry. Due to its shortness poetry is much more difficult to be analyzed quantitatively. Agglutinative languages such as Polish or Latin seem also to be a problem for Burrows's Delta.[23] Still under discussion is whether and if so which words should be removed (e.g. personal pronouns),[24] which class of words yields the highest accuracy[25] and at which level of most frequent words one gets better results,[26] the first hundred or the first 200 up to 300 most frequent words? To sum up Burrows's delta has become in its short time of existence well established, but still a lot of language specific research has to be done to understand its strengths and weaknesses in detail.

Using Delta for research has been helped a lot by the implementation in an R script which has been developed up to now by Maciej Eder and Jan Rybicki, who have lately been joined by Mike Kestemont.[27] Burrows himself has very early pointed out that stylistic measurements could be used in other fields of literary study outside of authorship attribution and he has done some research on using it on literary epochs.[28] There are other algorithms, scripts, and tools but Eder and Rybicki's scripts are working on the open source statics program 'R' which offers scalable opportunities to run analysis on small as well as on big (text) data. Moreover it is unique, as it is open for different delta analysis (Classic Delta, Argamon's Delta, Eder's Delta, Manhattan, Canberra, Euclidean),[29] and enables the user to adjust the best most frequent words parameters. It comfortably allows a wide range of analytical combination of style-markers settings, like 'culling' (i.e. culling rate specifies the percentage of texts in a corpus in which a given word must be found in order to be included in the analysis), the number of the Most Frequent Words (MFW) analyzed, deleting/non-deleting pronouns etc. The combination of a relatively simple statistical measurement and the existence of a tool which makes it easy to apply it on different texts, there are good reasons for us as non-experts in the field of language statistics to text it in the view of our knowledge of literary history.

In the following we will explore the possibilities this intellectual tool has for working with larger collections of German texts. Our goal is to evaluate Delta and its use for German literary history. Thereby we don't expect any new insights but want to evaluate the method in reference to the knowledge we already have. On the other hand if we can confirm the knowledge we have based on hermeneutic methods we succeeded in putting this knowledge on a new basis. We will approach this from two different angles. In the first section we will use Delta in a row of different tests. For example to group texts around 1800 according to authorship, epoch, genre, and gender. In the second section we will look at a specific problem of individual writing, the position of Kleist in literary history, and see what this instrument can contribute to the debate.

3. The Corpus

The basic goal of this section is to evaluate the usefulness of Burrows's Delta for the study of German literary history by applying it to a set of tasks like author attribution or classification based on genre or time or gender. As we know beforehand what should be the results of our measurements we can evaluate our tool by comparing the real outcome to the expected results. For more than one reason this can only be a first exploration of this kind of evaluation. The first and main one is a pragmatic one: the whole work is based on the validity of our corpus and its metadata. Because the text collection is so huge we concentrated first on the novels. They provide a larger chunk of text which makes the use of Delta more reliable.[30] And they are relatively few, only about 430, compared to the thousands of poems.

The texts have been part of one of the first large scale digitization projects in German language and most of them are based on scholarly editions which have been in academic use. On the positive side this means the texts are acceptable for scholarly work but on the other hand most of them have been published in a time when it was considered good practice to modernize the spelling. So the corpus is probably not of interest for anyone doing research on the history of spelling or other features depending on the exact form of the text surface. On the other hand this has solved or at least reduced a problem which is a big burden for all historical language research: the vast variety of spellings in earlier writing systems - in Germany before 1800. Or to put it the other way round, the modernization has in a kind done what otherwise would have to be done by the editors: the normalization of writing.

The corpus is not balanced or representative for the literary production in the time it covers. The main reason for this is its genesis: It has been produced by a commercial company which wanted to sell a DVD with all canonical literary texts and another DVD with literary texts by women. So women are overrepresented in this corpus compared to their share of the canon and probably also in relation to their share in literary production. The collection also contains translations and it has been shown, for example by Rybicki,[31] that stylometry can be used with interesting results in translation studies. But the translations concentrate on a very small canon and, because the contents of another DVD with erotic texts went into the collection, not a few of the translation in the 18th century belong more to the canon of erotic literature than to that of world literature. This shouldn't imply that only canonized works should be studied but rather point out that probably the text collection at hand is in some aspects especially unlikely to be representative. Another problem are the metadata for our research. Dates of publication were mostly missing in the texts and the same is true for the other parameters like epoch, the fact whether a novel is a translation, or gender.

Not enough with these problems of building a corpus: the genesis of extensive works like novels is often long and complicated. But this huge variety of different genetical processes cannot be captured easily in the Spartanian metadata of a corpus. For practical reasons we just used one field which contains the date of the first publication of the first volume or the year of the first number of the first publication in a journal, ignoring that many novels have been published over a range of years and that sometimes there has been a considerable time gap between the writing and the publication of a novel. We kept to this rule even in the case of posthumously published texts. Translations proved to be even more difficult, because we couldn't easily determine the time of their first publication. Therefore we determined to hold on preparing these metadata and just used the publication date of the original. This severely limits our possibilities to do anything interesting with the translations for now and will be one of the first steps to improve the quality of the corpus.

4. Results

This in mind let us start with a simple experiment: we look at all 63 novels in the corpus between 1785 and 1815, using the consensus tree based on about 27 iterations.

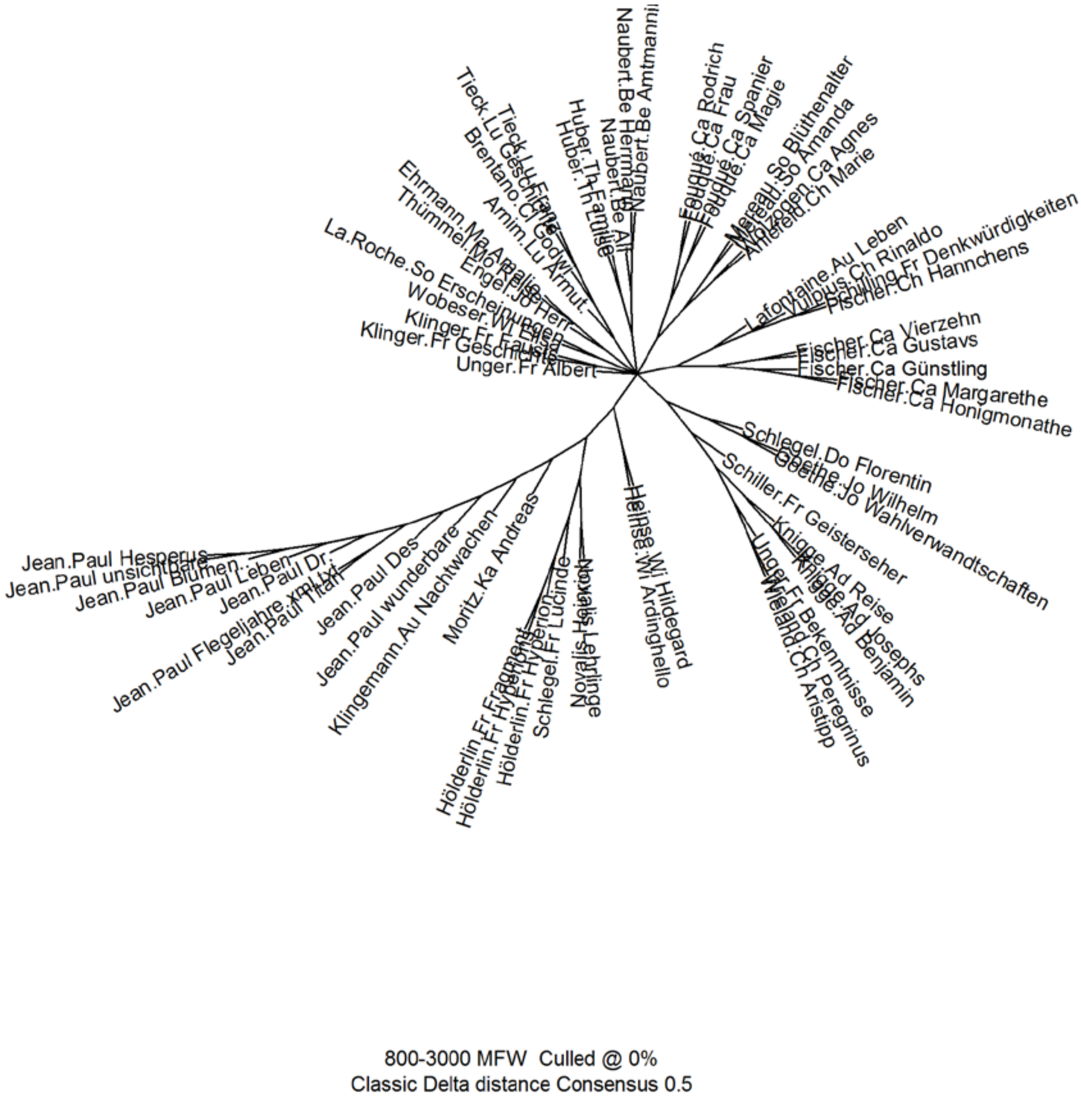

Fig. 1 “Novels between 1785-1815”: Consensus tree, Classic Delta, 800-3000 MFW, consensus 0.5

We can see two sections in the image: one the long branch containing Jean Paul’s novels mainly and the wheel at the top. This wheel shows 11 other branches all starting from the center which basically means that you cannot say anything about the stylistic distance between these branches. Only the length of the lines represents the distance between the texts. One point is rather obvious: Almost all authors are correctly assembled into sub branches; in other words, Delta is indeed a very good indicator for authorship for this corpus. There seems to be one error: Friederike Helene Unger’s novel *Bekenntnisse einer schönen Seele (Confessions of a beautiful soul)* from 1806 isn’t put together with her novel *Albert und Albertine*. But actually this could be not an error at all but an insight. *Bekenntnisse* has been published anonymously and it is anything but clear whether the novel really is written by Unger. There are two other writers mentioned as possible authors of the book: Paul Ferdinand Buchholz and Charlotte von Ahlefeld.[32] If we focus on a smaller group of texts (classic Delta, 3000 MWF) including also another novel by Unger and other female authors but also the novels by Tieck and Goethe, we still see that *Bekenntnisse* is not grouped together with Unger’s other novels:

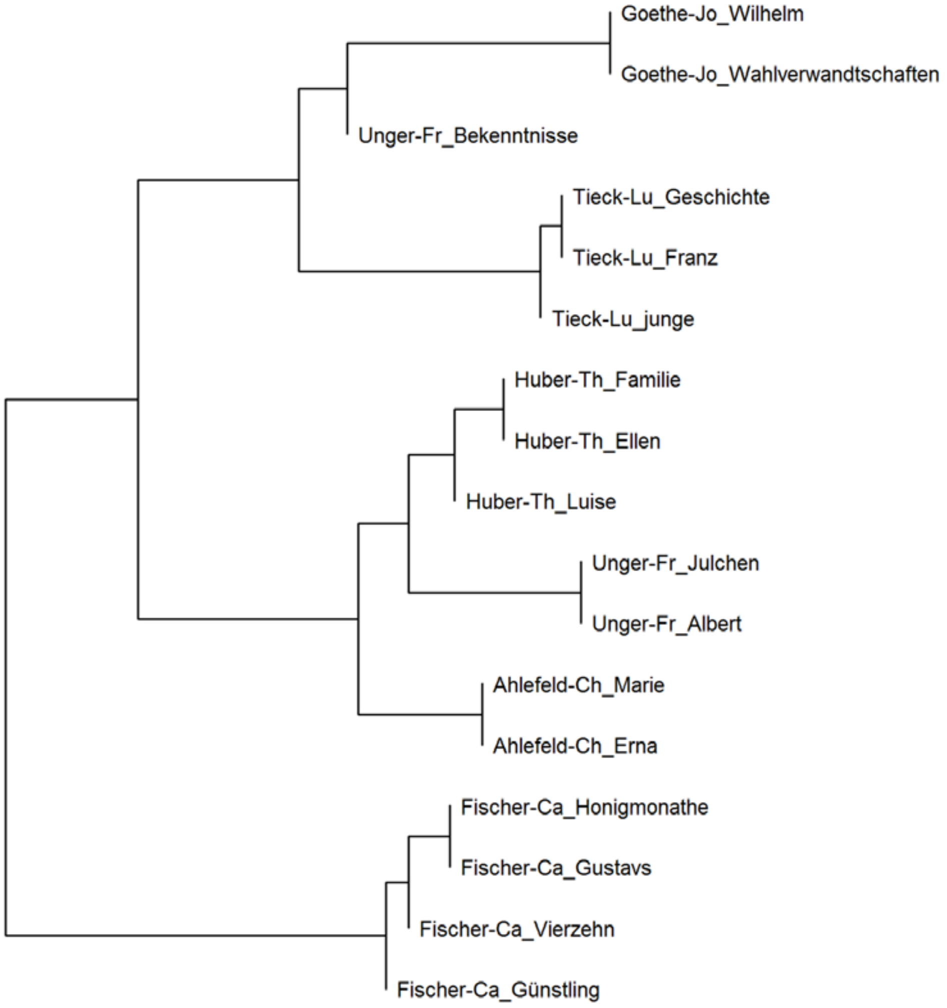

Fig. 2 “Therese Unger, *Bekenntnisse einer schönen Seele* among the novels”: Dendrogram, Classic Delta, 3000 MFW

Neither is Unger grouped together with the novels of Charlotte v. Ahlefeld nor with Fischer or Huber. This result seems to be relatively robust with different settings and different additions of other authors, so probably it would be worthwhile to have a closer look at the third candidate, Buchholz. But as there are none of his texts are available in digital form at the moment, we cannot pursue this line of inquiry any further.

It is a well established fact in stylometry that genre is one of most important aspects of style. In other words, one can expect, that genre should be a good classifying feature. As we already have seen authorship is a very good classifying feature, so in the following run we keep this dimension invariant. In the following test, we use a collection of 39 texts from different genres but all written by the same author, by Goethe. The prefix marks the genre, based on our conventional understanding of genre. As figure 3 shows the grouping works rather well but includes some interesting unexpected results:

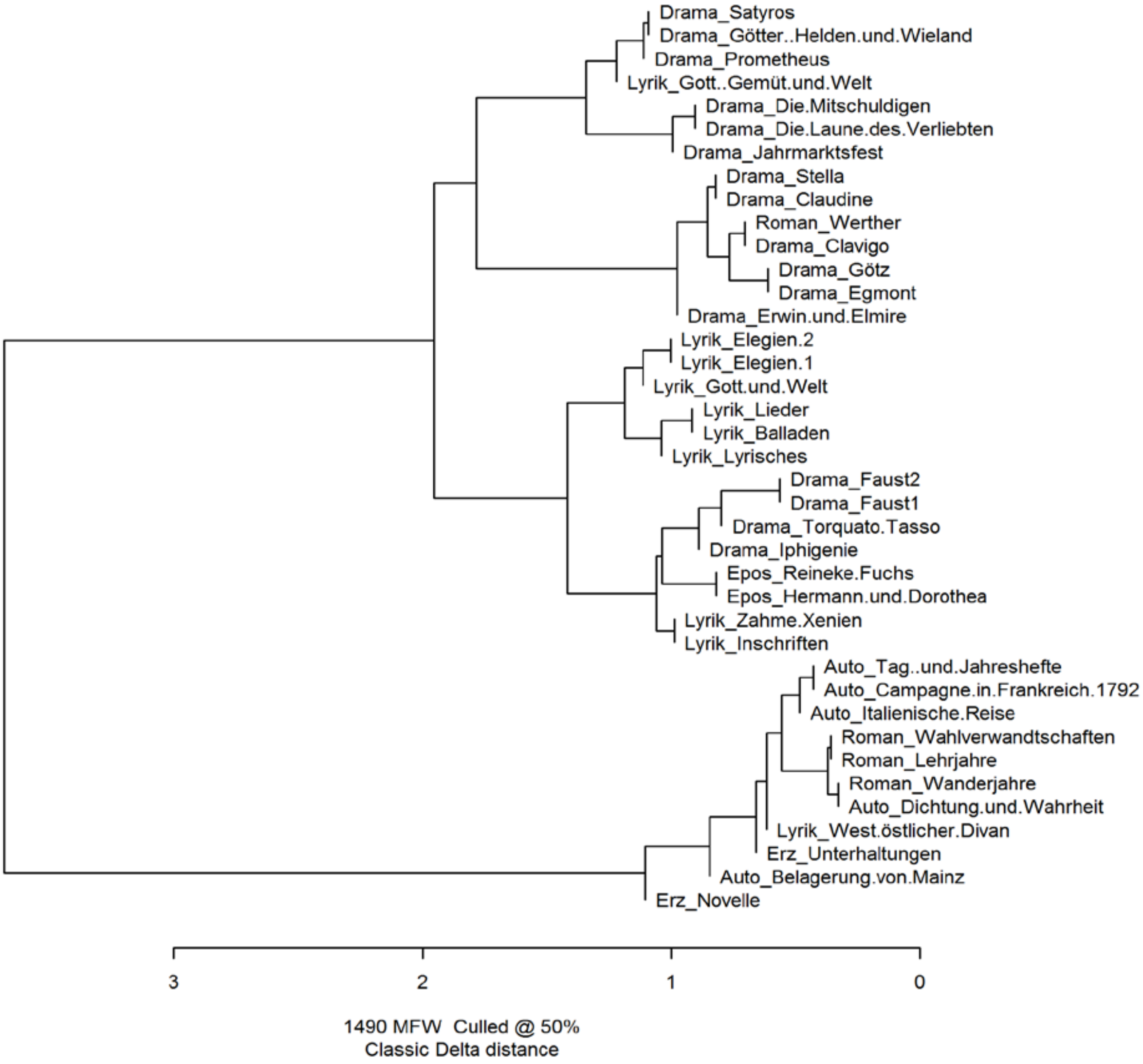

Fig. 3 “Goethe’s work, grouped according to genre”: Dendrogram, Classic Delta, 1490 MFW, culling 50%

There are two distinct groups. One consists of four subgroups and the other of the one group shown at the bottom. The group at the bottom contains all prose texts, the novels, the autobiographical texts and the smaller novellas. The four groups at the top contain the lyrical texts, the epos and the plays. Group 1 and 2 (counting from the top), with texts stylistically closer to each other, contain with only one exception plays, while group 3 only contains lyrical texts, while group 4 is a mix of 4 plays, 2 Epos and 2 collections of lyrical texts, and interestingly enough they are grouped internally along these genre divisions. But obviously also time is a factor here: the plays in the top group have mostly been written by the young Goethe or in the early years of his stay in Weimar, while *Tasso*, *Faust* (at least in this version), *Iphigenie* are from his classical period and the poems collected in *Zahme Xenien* have been written partly soon after 1815 and between 1824-27. And the *Inschriften, Denk- und Sendeblätter* have also written during these late years.[33] His early novel, *Die Leiden des jungen Werther*, is an interesting outlier, because it is the only prose text where the genre attribution didn't work. This can be easily explained by the fact that it is an epistolary novel consisting only of one voice, that of Werther. But the 'false' grouping could also be understood as a hint to look closer at the affinity between the monologue in plays and this novel. All in all we have seen that genre is as important as expected and is an important factor in clustering texts using a stylistic measurement like Delta, but it is not as reliable as authorship--which could be seen as an ironical comment on the debate on authorship following Barthes' and Foucault's essays.[34]

A quantitative approach is not only a method to tell authorship apart. Distant reading is also a tool to see through literary history. Here we are comparing two test sets of authors who are regarded as typical of their respective times (Enlightenment and Realism) and it is astonishing enough that the 24 novels cluster so neatly along the group boundaries. But maybe this is an effect of the words used. So what happens, when we ask the program to use only those words which are common to all texts? In the dendrogram above the horizontal axis maps the edges, that is the distance between the texts which are historically nearly hundred years away from another.

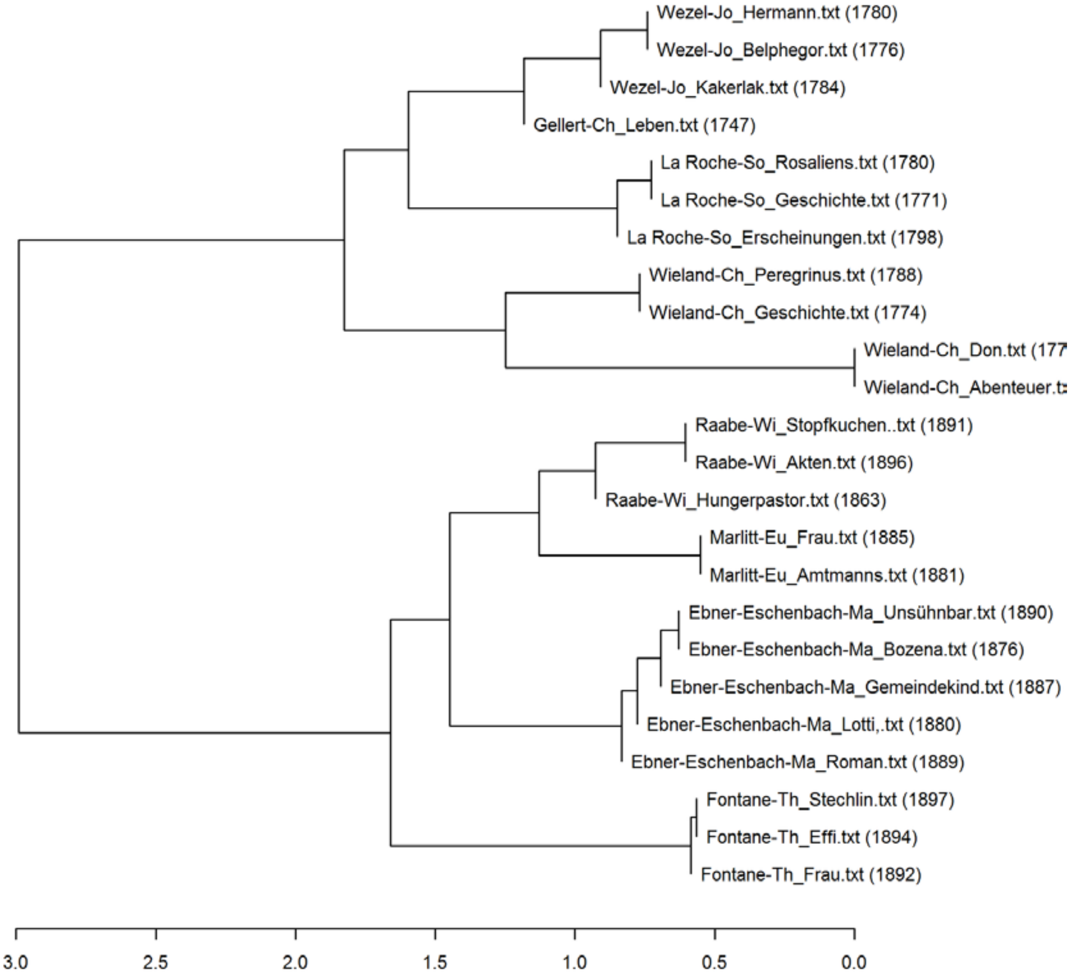

Fig. 4 “Distinction between literary epoch: Enlightenment vs. Realism”: Dendrogram, Classic Delta, 2000 MFW

The stylometric approach shows a clear cut division between the texts of Enlightenment and of Realism. Later on we will discuss examples where the two epochs are not so far apart, but maybe it is worthwhile to point out that a simple procedure to classify unknown texts reliably into a specific epoch or time span would be a very valuable tool for handling very large corpora.

There is more to it than that if we take books into account which were often left beside by literary history or put into a separate collection like the writing of woman. Here we choose 20 novels written by female authors and 17 written by men to see how these German novels around 1800 are grouped by their stylistic features. We decided to delete the pronouns to avoid the influence of having main characters with different gender though this could regard as an interesting distinguishing feature too.

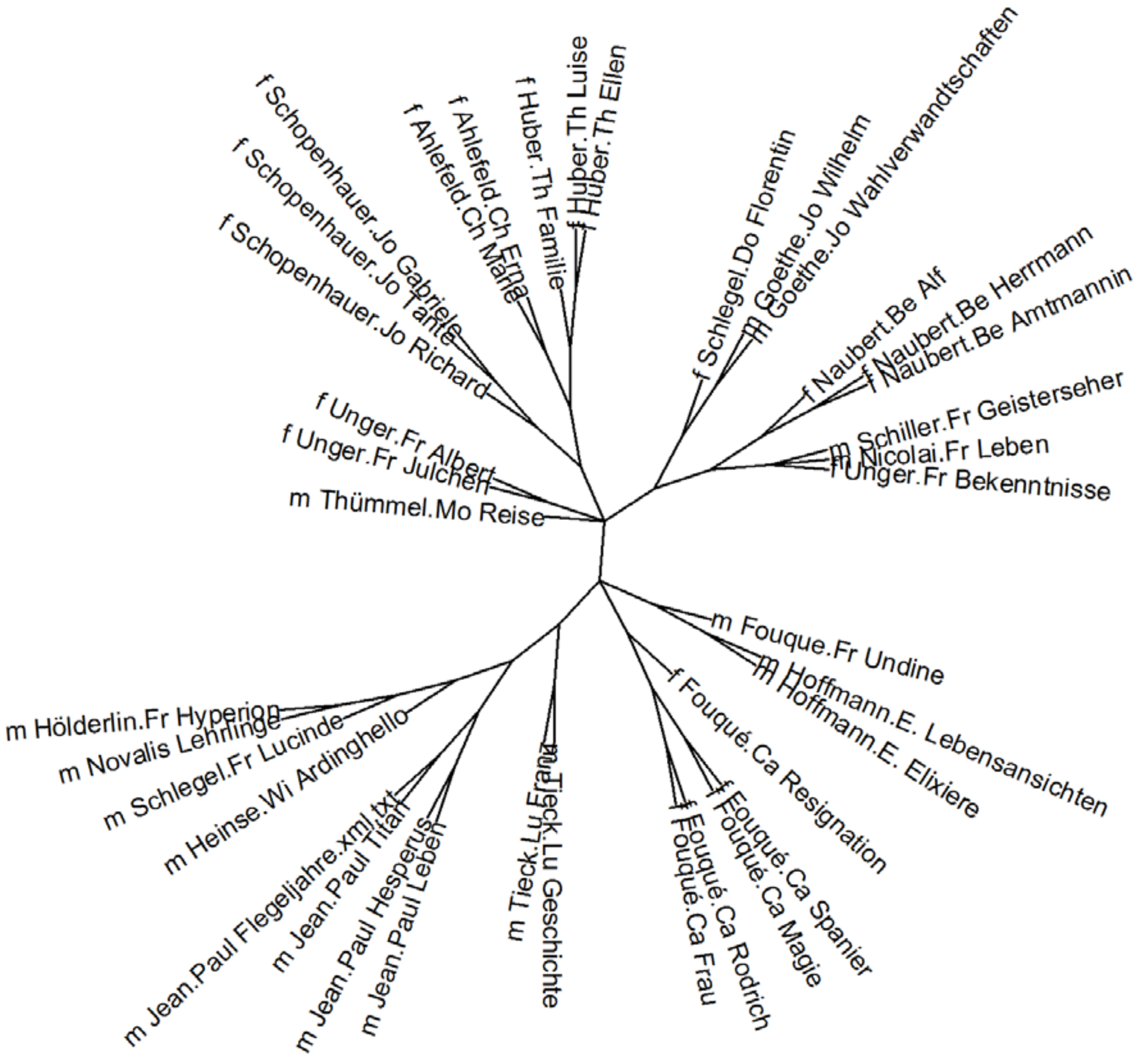

Fig. 5 “Male and female authors around 1800”: Dendrogram, Classic Delta, 200-2500 MFW, consensus 0.5, pronouns deleted

Female and male authors cluster mostly separately from one another, but this could be a mere effect of good authorship attribution. (And again *Bekenntnisse einer schönen Seele* shows a noticeable distance to Friederike Helene Unger's writings). Though we also found some evidence to the contrary we couldn't classify reasonably well male from female texts – the results were very dependent on the setup of the group as a whole. The really interesting thing here is the literary landscape around 1800 which becomes visible through this mapping out of stylistic affinities. Charlotte v. Ahlefeld, Therese Huber and Johanna Schopenhauer form a literary province of female writing. But other female writers like Caroline de la Motte Fouqué who belongs to the romantic camp or Benedikte Naubert are clustered into the vicinity of male authors. There seems to be no general stylistic feature of female writing but a limited number of positions in the literary field around 1800. Some of these positions are held exclusively by female authors, some like Dorothea Schlegel write similar to Goethe, some like Caroline de la Motte Fouqué do not belong to the female position in the literary field of that time. This could correspond to findings of gender studies where the bias in constructing the canon and the continous disregard of female traditions has been has been pointed out.[35] With findings like this distant reading opens the view on the structure of a historical literary field which is more than the replication of the canonical picture. It seems to us that this constellation has to be explored more in the future, even though we did experience too that some of the classifications were all but stable and seemed to be especially dependent on the settings of the parameters.

In a second series of tests we analyze the historical position of an individual author. Heinrich von Kleist’s work oscillates between classicism and romanticism and literary history widely debates where to put him. Therefore he is a good example whether stylometry can contribute arguments for a better historical indexing of Kleist. For a first pretest we took a smaller corpus of 32 highly canonical works of German dramatic literature around 1800, but together with some popular dramas by Iffland and Kotzebue. Again the texts, that constitute the corpus, are plain text files. And again such a corpus deals with a bunch of problems in details. The number of word types differs widely. Dramas present the speech of roles which is not comparable with the voice of a narrator. And especially the longstanding tradition to distinguish comedy and tragedy might be a hint that drama is not a genre but two. Despite these problems we put them into brackets and used 2000 most frequent words analysis, without culling or deleting pronouns nor using any list of stop word. The distance is measured again by Classic Delta.

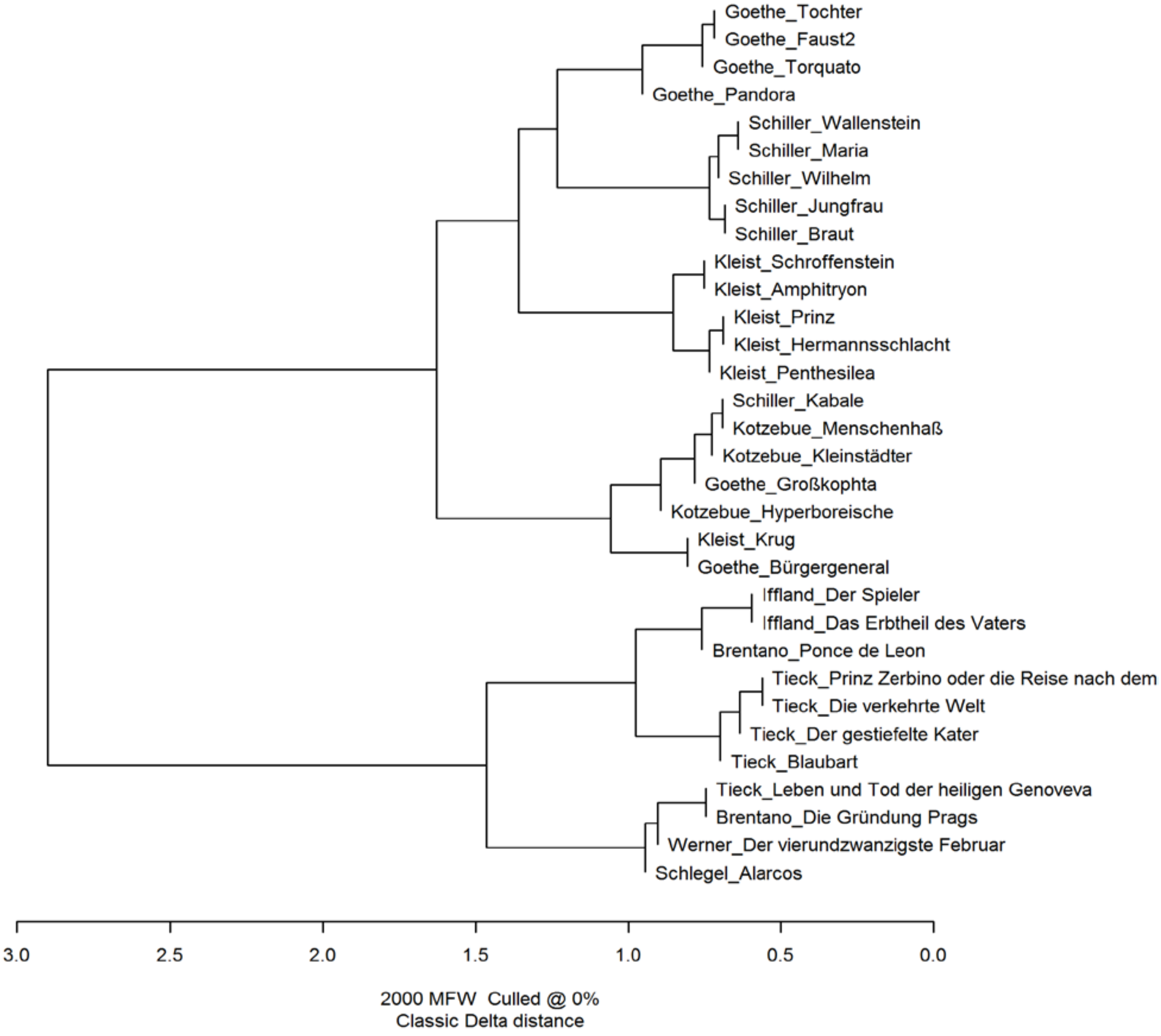

Fig. 6 “Kleist’s drama among other dramas around 1800”: Dendrogram, Classic Delta, 2000 MFW

As usual the distance between the dramas is represented by the horizontal length of the lines, while the vertical distance is not of any importance. The first results gave a cue: Kleist is rather close to classical dramatists like Goethe and Schiller and in an obvious distance to the group of romantic plays. Interestingly enough we can see in the group between classical and romantic plays, but closer to the classical texts, a wild mixture of texts by Goethe, Kotzebue and Kleist, which can be explained by the fact that all these texts are comedies and it seems that the specific language of comedies marks them even more than their author does. The classification of Schiller's *Kabale and Liebe [Intrigue and Love]* into the context of comedy seems to be a fluke at the moment if it is not caused by the more comical parts of this play. Probably Kleist's *Der zerbrochene Krug [The Broken Jug]* is right to be closer to the domestic tragedy genre than Kleist's other comedy *Amphitryon*.

This is only a first hint how methods of distant reading could catch Kleist's characteristic style with only formal style-markers. To make more than a small point we go a step further and take a list of 49 dramas written or published between 1790-1811. Again we use the scripts by Eder and Rybicki but now we go from 200 to 2000 MFW with a consensus strength of 0.5, which means that 50% similarity between the single dendrograms behind the consensus tree should be given to be use as a characteristic to measure how similar two texts are.

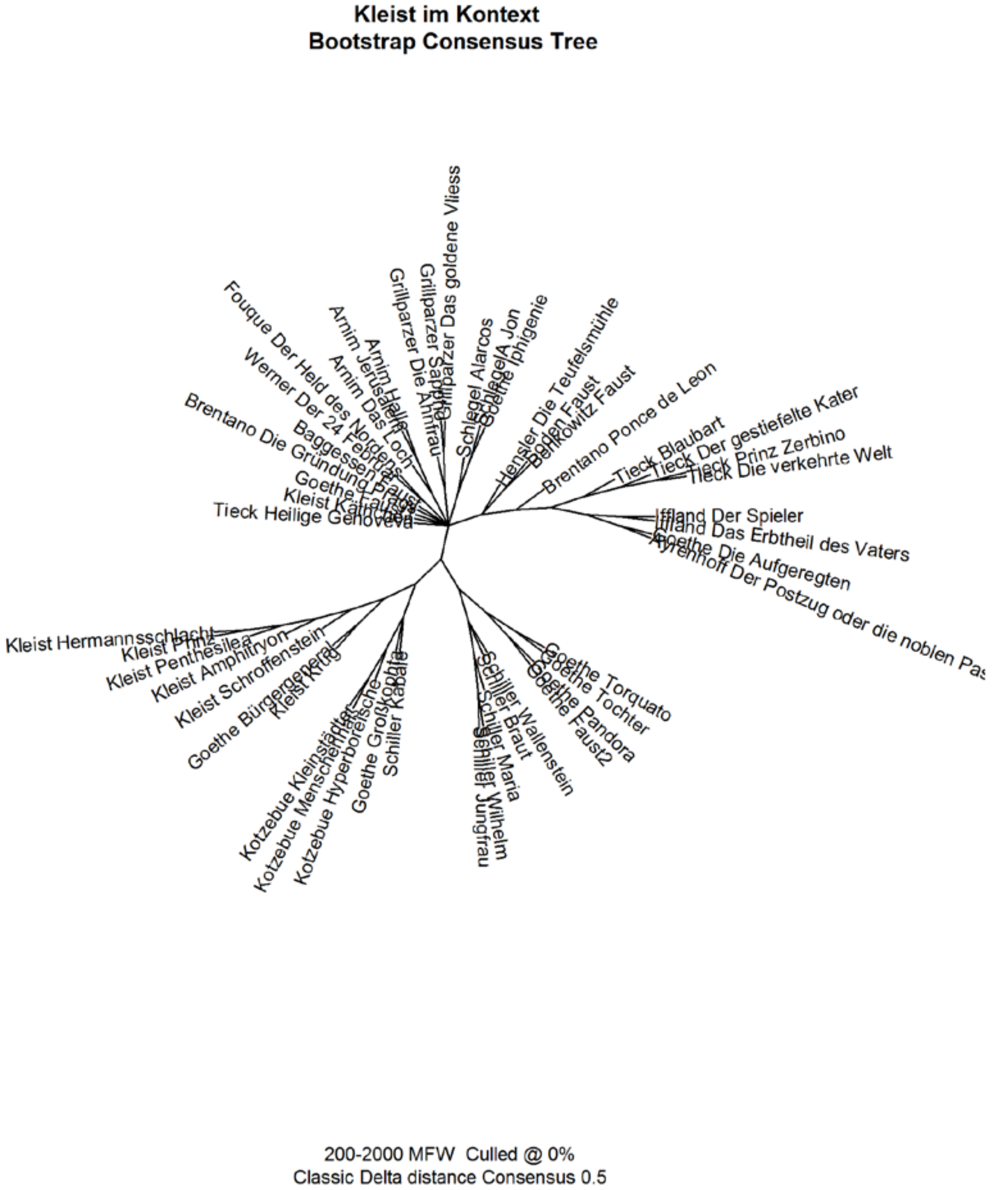

Fig. 7 “Kleist’s dramas and other dramas around 1800”: Consensus tree, Classic delta, 200-2000 MFW,

In contrast to the dendrograms the figure here shows that Kleist’s dramas cluster closely with one exception, that is his *Käthchen von Heilbronn*. By the use of other deltas like that of Eder or Argamon’s Delta *Käthchen* stays always apart from Kleist’s other dramas and is closer to the romantic dramas. This is consistent with the special genre Kleist has used for this drama. He called it “ein großes historisches Ritterschauspiel [a big historical knights play].”[36] The branch also retraces the development of Kleist’s writing of tragedies, starting with his first tragedy *Die Familie Schroffenstein* (1803) and ends with the very similar *Hermannsschlacht* (1808) and *Prinz von Homburg* (1809). Analog the algorithm maps the dramatic work of Schiller.

To avoid any form of so called cherry picking, that is in order not to suppress evidence, which contradicts the expected results, we have to scrutinize more closely how important deleting/non deleting of pronouns/culling and the chosen delta algorithm are. With culling the dramas are more similar to each other (a) than without (b) and seem to lose some of their specific stylistic features, here endorse by deleting pronouns. By culling the algorithm seems to throw out too many of those words which distinguish the romantic characters of *Käthchen* from Kleist’s other characters.

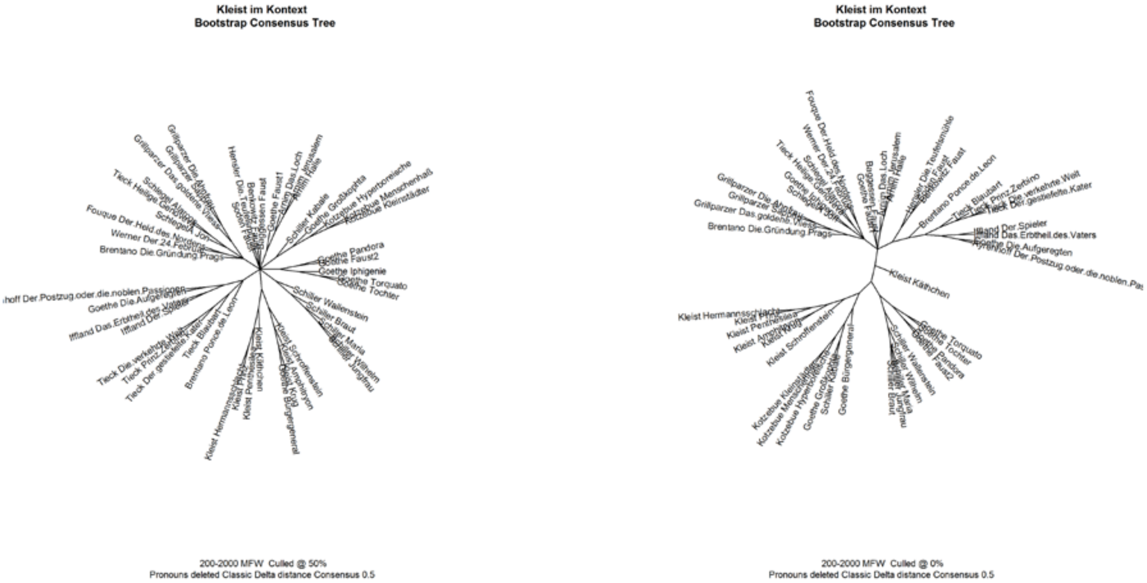

Fig. 8 “Kleist’s dramas and other dramas around 1800”: Classic Delta, 200-2000 MFW, pronouns deleted, consensus 0.5. a) culling 0.5, b) culling 0.0

The chosen delta also plots slightly different results. Only slightly because no matter whether we choose Classic Delta or Eder’s Delta except of *Käthchen* Kleist’s dramas are always nearby themselves and always in a wide difference to the romantic dramas of his time.

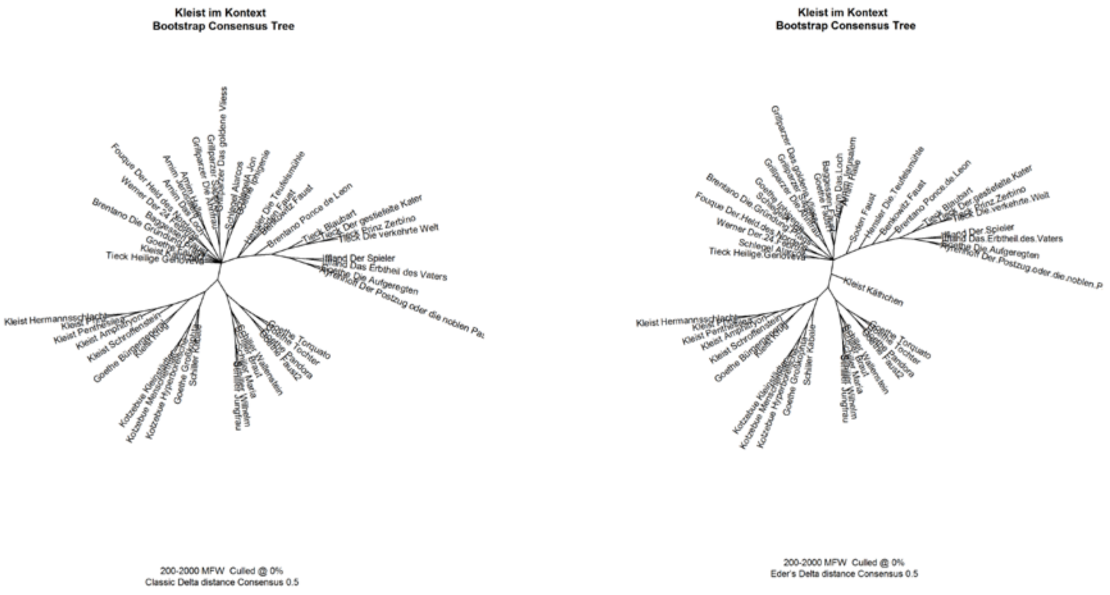

Fig. 9 “Kleist’s dramas and other dramas around 1800”: 200-2000 MFW, consensus 0.: a) Classic Delta, b) Eder’s Delta

If we compare these consensus trees with the dendrogram of the beginning of our second series of tests together they show a robust finding. Kleist’s style of drama separates him from the romantic writing of his time very clearly. Only his *Käthchen*-drama is more closely to the romantic plays. A difference between dendrograms and consensus tress is how close Kleist’s dramas are in comparison with the domestic tragedies. Another difference between a 2000 MFW and the 200-2000 MFW algorithm is the identification of genre. The dendrograms find better the differences between comedy and tragedy while the consensus tree is better in plotting the author’s style. Further research is needed whether drama in general plots more diverse results than novel does. Around 1800 it might be highly plausible that drama was much more classified in many sub-genres while novel is an emerging genre with less sub-genres.

Our last arrangement is a bit unusual even in this yet not so common setting. As we have pointed out above the stylistic affinities between texts could be interpreted, especially if not only one text of the author is concerned but all or at least most of her or his novels, as a specific view on the literary landscape at this span of time. There are other instruments at hand to reconstruct such a landscape, but usually we use social factors (as in a literary field analysis according to Pierre Bourdieu) or mix of different factors as is usual in traditional literary history. So the question seems to be how can we relate the clusterings, the groupings, the stylistic affinities which seems to be suggested by the stylistic analysis to those mapped by these other methods. Is it possible to generalize our findings for specific works to the work of an author in general? One approach we tested was to cumulate all novels by one author into one file and to look at the results of our Delta procedure:

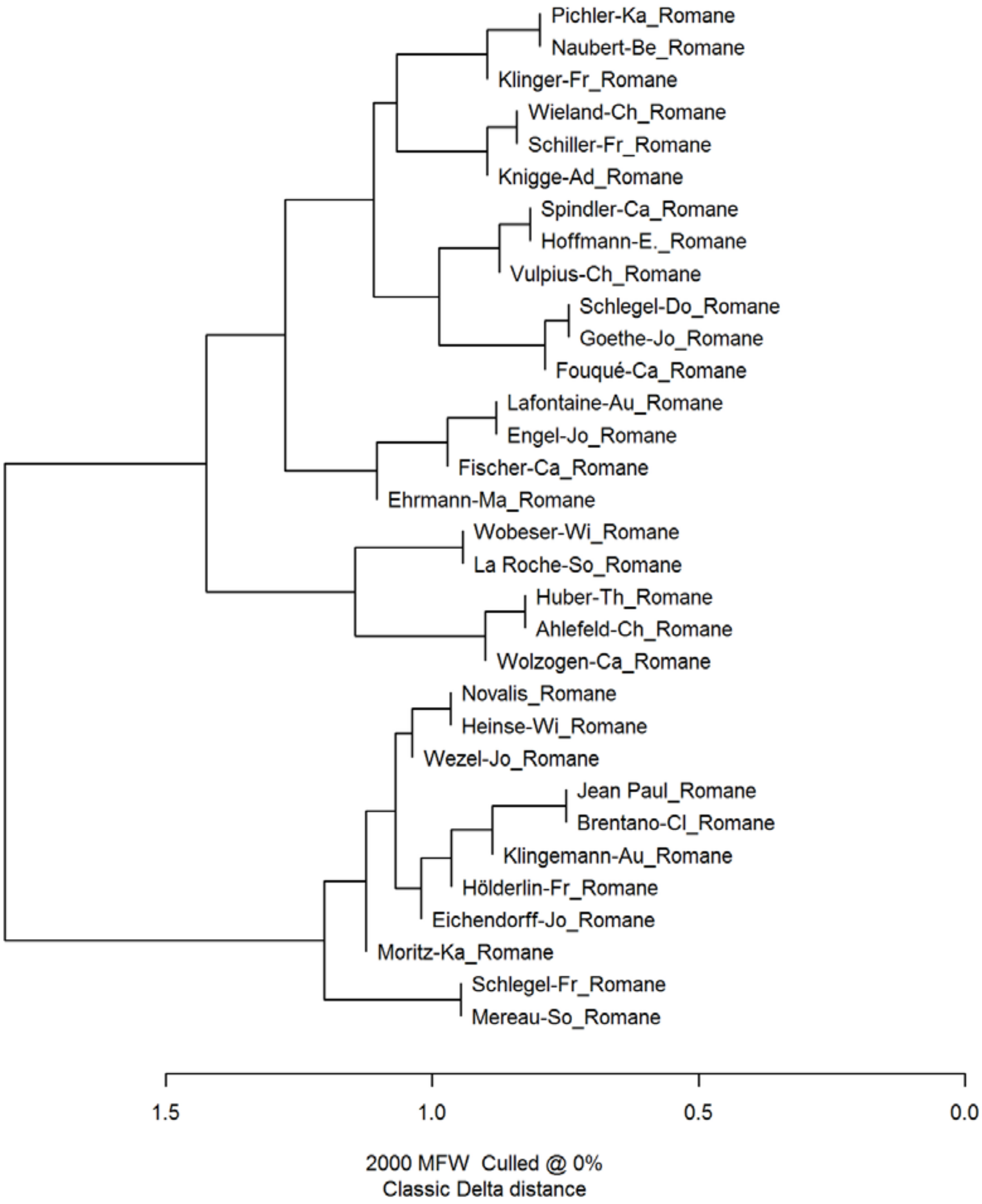

Fig. 10 “Authors of novels as a literary field around 1800”: Classic delta, 2000 MFW

It is easy to see that there are two groups: the first one is the larger group above with Goethe, Schiller, Wieland and many others, while the smaller one below contains many names of writers usually classified as belonging to the era of Romanticism like Eichendorff, Schlegel, Brentano, Tieck. Interestingly enough two outsiders of literary history, Jean Paul and Hölderlin, are part of this group too as are writers strongly associated with the Enlightenment like Wezel. And again we can discern a group of female writers (the names directly above Novalis) who haven't been really been perceived as belonging together, except for the fact that they are women. But this quality they share with women who do not belong to this group, for example Dorothea Schlegel who in all runs we did always found herself next to Goethe. So maybe this impression of a group of mostly female writers may be worth another visit to the literary history of the time around 1800 in order to find out whether these stylistic affinities correspond to other feature too.

5. Discussion

Stylometry is an example for distant reading. It shows how well quantitative analysis works also for research in (German) literary history. It is quite capable of telling author from author, Jean Paul or Kleist from other authors around 1800 and could discriminate genres and gender, epochs and authors. Eder's and Rybicki's scripts offer style marker settings which transform quantitative text analysis from an arcane knowledge into a practical tool. For the first time a quantitative approach could work on large corpora and we expect that literary historians will appreciate the new research possibilities as soon as they jump over the two culture gap.

But stylometry is not a science machine where you could pose your questions and the script coughed up the results. Carefully series of tests and a deep knowledge of the analyzed texts are necessary precondition. Otherwise distant reading would do not more than cherry picking. It has to keep in mind how strongly a better understanding of statistics would improve distant reading. There are good arguments why statistics should become part of (literary) scholarly education in the next years.

Burrows's delta is useful for literary history but seems to work with different validation on different texts and genre. Authorship attribution is what works best also in our test series. Genres sometimes make a difference but it depends on the genre and his historical types of sub-classifications. Discrimination of epoch also works well within the limits of genre, here in our research with novels. Still the hereby used corpora are not sufficient to answer questions on sub-genres precise enough. Similar is the effect of missing authors. The pure lack of non-canonical but culturally significant texts like that of Buchholz limits the extent of our conclusions more than once. A major task for further research is therefore the procurement of better and larger corpora. The test corpora used here are not historically valid enough as far as they do not map the written, printed and read books of their time precisely enough. What they represent is a more or less random sample of books canonized through the history of literature, preference of today's publishing house, scholarship and literary critic. A culturally, historical, and social significant corpus looks differently. If we want to find out whether the group of mostly female writers really represents a stylistic position in the literary field around 1800 next to the canonized positions of classical and romantic authors, if we want to decide whether Unger's *Bekenntnisse einer schönen Seele* are written by Ferdinand Buchholz, we simply need more texts, texts by other male and female writers and of authors like Buchholz. For a representative corpus of the literature around 1800 it would be a precondition that one could work on a representative corpus of texts at a certain period of time. Once more it become obvious that quantitative research needs a qualitative understanding about a certain point of literary history – and the other way round. Otherwise no representative corpus could be build.

One of the most interesting point of Burrows's Delta is his use of every word. Contrary to the many approaches that make use of stopword lists Burrows's, Eder's and Rybicki's stylometry take care of every little word. There are good arguments by psycholinguistic research how significant these seemingly innocuous words are for the individual style. As James Pennebaker and others have shown function words like pronouns, articles, prepositions, auxiliary verbs serve as a good indicator of personal style.[37] Burrowsian stylometry, though, makes not exclusively use of them, it utilize widely these most common words. It would be important to know more about the link between psychology, word use and stylometry. This could give explanations for the deeper mechanism behind the delta.

Quantitative text analysis is scalable reading because its method and a methodology catch individual style like that of Jean Paul or Heinrich von Kleist as well as groups or epochs of writings styles. It offers not only the validation of already known literary constellations but give a wider picture of the literary field. Which works cluster to a position, how many of those positions are in the field under specific historical conditions, how large or how small are these positions—questions like this could be better answered by integrating quantitative research methods into literary history. And most frequent words are not the only features one could analyze. Sentiment analysis, genre specific features like narrative or dramatic style could serve as further settings in a multivariant quantitative analysis of literary history. And to compare the results with findings of other European and maybe someday with other Non-

European text analysis is more than a little research program. There is more than the canon and distant reading is a road to go beyond.

[1] Hermann Hettner, *Literaturgeschichte des 18. Jahrhunderts*, vol. 3 (Braunschweig. Vieweg, 1856-70), 526-541; David Wellbery (Ed.), *A New History of German Literature* (Harvard: UP, 2005), 511-516 jumps from Hölderlin 1808 directly to Brother Grimm 1815. Fair to say that (for history of German literature) starting from Wilhelm Scherers *Geschichte der deutschen Literatur*, 1883, over Leo Balet / E. Gerhard, *Die Verbürgerlichung der deutschen Kunst, Literatur und Musik im 18. Jahrhundert*, 1936, to the social history of literature like *Hansers Sozialgeschichte der deutschen Literatur. Vol. 3: Deutsche Aufklärung bis zur Französischen Revolution 1680-1789*, ed. Rolf Grimminger, 1980, there was always a wider knowledge about the history of books. But labels like positivism, sociology of literature, history of books, history of reading etc. name a niche in current literary studies.

[2] Christine Haug, "Buchserien und Anthologien. Wirkungsmächtige Medien zur Etablierung und Durchsetzung von ausländischen Literaturen in Deutschland im 19. Jahrhundert," *IASL-Online* (10.09.2005), URL: <http://www.iaslonline.lmu.de/index.php?vorgang_id=2122> (10.9.2012).

[3] Barbara Kastner, "Statistik und Topographie des Verlagswesens," *Geschichte des deutschen Buchhandels im 19. und 20. Jahrhundert*, ed. Georg Jäger, vol. 1/2 (Frankfurt/M.: MVB, 2003), 300-367; cf. also Ilsedore Rarisch, *Industrialisierung und Literatur. Buchproduktion, Verlagswesen und Buchhandel in Deutschland im 19. Jahrhundert in ihrem statistischen Zusammenhang* (Berlin: Spiess, 1976); *Publishing Culture and the "Reading Nation": German Book History in the Long 19th Century*, ed. Lynne Tatlock (Rochester: Camden, 2010).

[4] Gregory Crane: "What do you do with a million books? ," *D-Lib Magazine* 12, 6 (2006), http://www.dlib.org/dlib/march06/crane/03crane.html.

[5] URL: <www.textgridrep.de> (10.9.2012).

[6] Leonid Taycher, "Books of the world, stand up and be counted! All 129,864,880 of you," *Google Books Search*, URL: <http://booksearch.blogspot.de/2010/08/books-of-world-stand-up-and-be-counted.html> (10.09.2012).

[7] Google books ngram viewer: <www.books.google.com/ngrams> (10.9.2012).

[8] See for example Dan Cohens posting „Initial Thoughts on the Google Books Ngram Viewer and Datasets“ <http://www.dancohen.org/2010/12/19/initial-thoughts-on-the-google-books-ngram-viewer-and-datasets/> (11.11.2012)

[9] Franco Moretti, “Conjectures on World Literature,” *New Left Review* 1 (2000), 54-68, URL: <http://newleftreview.org/II/1/franco-moretti-conjectures-on-world-literature> (10.9.2012).

[10] Stanley Fish, “Mind Your P’s and B’s: The Digital Humanities and Interpretation,” *New York Times* (23.1.2012), URL: <http://opinionator.blogs.nytimes.com/2012/01/23/mind-your-ps-and-bs-the-digital-humanities-and-interpretation/> (10.9.2012).

[11] Martin Mueller, “Stanley Fish and the Digital Humanities,” *Center for Scholarly Communication & Digital Curation Blog*, URL: <http://cscdc.northwestern.edu/blog/?p=332> (10.9.2012), cf. also his Blog *Scalable Reading* (2/8/2012), URL: <http://scalablereading.org> (10.9.2012).

[12] Cf. Jean-Baptiste Michel et al., “Quantitative Analysis of Culture Using Millions of Digitized Books,” *Science* 331, 6014 (2010), 176-182.

[13] Russel Horton et al., “Mining Eighteenth Century Ontologies. Machine Learning and Knowledge Classification in the *Encyclopédie*,” *Digital Humanities Quarterly* 3, 2 (2009), <http://www.digitalhumanities.org/dhq/vol/3/2/000044/000044.html> (10.9.2012).

[14] Clay Templeton, “Topic Modeling in the Humanities: An Overview,” *Maryland Institute for Technology in the Humanities Blog*. posted 1 August 2011, <http://mith.umd.edu/topic-modeling-in-the-humanities-an-overview/> (10.9.2012).

[15] Franco Moretti, “Network Theory, Plot Analysis,” *Stanford Literary Lab Pamphlets* 2 (01.05.2011), <http://litlab.stanford.edu /LiteraryLabPamphlet2.pdf> (10.9.2012).

[16] We follow *Contributions to the Theory of Text and Language. Word Length Studies and Related Issues*, ed. Peter Grzybek (Dordrecht: Springer, 2006), 15-60.

[17] A good example is the work of the statisticians Frederick Mosteller & David L. Wallace, *Inference and Disputed Authorship: The Federalist*. (Reading/Mass.: Addison-Wesley, 1964). They use function words to determine authorship of the anonymously published *Federalist Papers*.

[18] Cf. David L. Holmes, “The Evolution of Stylometry in Humanities Scholarship,” *Literary & Linguistic Computing* 13, 3 (1998), 111-117.

[19] <http://voyant-tools.org> (10.9.2012).

[20] John Burrows, “Delta: A Measure for Stylistic Difference and A Guide to Likely Authorship,” *Literary & Linguistic Computing* 17, 3 (2002), 267-287.

[21] David L. Hoover, “Testing Burrows’s Delta,” *Literary & Linguistic Computing* 19, 4 (2004), 453-475.

[22] Karina van Dalen-Oskam & Joris van Zundert, “Delta for Middle Dutch—Author and Copyist Distinction in *Walewein*,” *Literary & Linguistic Computing* 22, 3 (2007), 345-362.

[23] Jan Rybicki & Maciej Eder, “Deeper Delta across genres and languages: do we really need the most frequent words?,” *Literary & Linguistic Computing* 26, 3 (2011), 315-321; Maciej Eder & Jan Rybicki, “Do birds of a feather really flock together, or how to choose training samples for authorship attribution,” *Literary & Linguistic Computing*, preprint online http://llc.oxfordjournals.org/content/early/2012/08/10/llc.fqs036.abstract.

[24] David L. Hoover, “Testing Burrows’s Delta,” *Literary & Linguistic Computing* 19, 4 (2004), 453-475.

[25] Marius Popescu & Liviu P. Dinu, “Comparing Statistical Similarity Measures for Stylistic Multivariate Analysis,” *Proceedings of the International Conference RANLP* (2009), 349-354.

[26] Peter W. H. Smith & W. Aldridge, “Improving Authorship Attribution: Optimizing Burrows‘ Delta Method,” *Journal of Quantitative Linguistics* 18, 1 (2011), 63-88.

[27] Maciej Eder & Jan Rybicki, “Computational stylistics,” <https://sites.google.com/site/computationalstylistics/> (10.9.2012). Their use ot this tool to a wide set of problems has been discussed in Jan Rybicki & Maciej Eder, “Deeper Delta across genres and languages: do we really need the most frequent words?,” *Literary & Linguistic Computing* 26, 3 (2011), 315-321 and other papers.

[28] John Burrows, “Questions of Authorship: Attribution and Beyond,” *Computers and the Humanities* 37, 1-26; John Burrows, “Textual Analysis,” *A Companion to Digital Humanities*, ed. Susan Schreibman, Ray Siemens, John Unsworth (Oxford: Blackwell, 2004), 323-347.

[29] Cf. for the mathematical background Shlomo Argamon, “Interpreting Burrows’s Delta: Geometric and Probabilistic Foundations,” *Literary & Linguistic Computing* 23, 2 (2008), 131-147.

[30] David L. Hoover, “Testing Burrows’s Delta,” *Literary & Linguistic Computing* 19, 4 (2004), 453-475.

[31] Magda Heydel & Jan Rybicki, “The Stylometry of Collaborative Translation,” <http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/the-stylometry-of-collaborative-translation/> (10.9.2012).

[32] Cf. Susanne Zantop in her afterword to [Friederike Helene Unger]: Bekenntnisse einer schönen Seele von ihr selbst geschrieben. Berlin 1806 [ND Olms 1991], p. 387f.

[33] See the commentary in Goethe, *Gedichte 1800-1832*, ed. Karl Eibl (Frankfurt/M.: Deutscher Klassiker Verlag, 1988).

[34] Actually John Burrows was one of the first to point out that Foucault’s claim that the ‘author’ is a mere projection of the reader is in contradiction to stylometric findings, but the common interest in his fact-based argument wasn’t really high; see John F. Burrows, “Computers and the Idea of Authorship,” *Rückkehr des Autors. Zur Erneuerung eines umstrittenen Begriffs*, ed. Fotis Jannidis, Gerhard Lauer, Matias Martinez, Simone Winko (Tübingen: Niemeyer, 1999), 167-182.

[35] Cf. the survey in Renate von Heydebrand / Simone Winko: Arbeit am Kanon. Geschlechterdifferenz in Rezeption und Wertung von Literatur. In: Hadumod Bußmann / Renate Hof: *Genus. Zur Geschlechterdifferenz in den Kulturwissenschaften* (Stuttgart: Kröner) 206-261.

[36] Heinrich von Kleist, *Das Käthchen von Heilbronn* (Berlin: Realschulbuchhandlung, 1810).

[37] James Pennebaker, *The secret life of pronouns. What our words say about us* (New York: Bloomsbury, 2011).