Quantitative Text Analysis for Literary History - Report on a DARIAH-DE Expert Workshop

Christof Schöch, Fotis Jannidis

2013

Quantitative Text Analysis for Literary History

Report on a DARIAH-DE Expert Workshop

Christof Schöch, Fotis Jannidis

DARIAH-DE Working Papers Nr. 2

Schöch, Christof ; Jannidis, Fotis: Quantitative text analysis for literary history : report on a DARIAH-DE expert workshop
Göttingen : GOEDOC, Dokumenten- und Publikationsserver der Georg-August-Universität, 2013
(DARIAH-DE working papers 2)

Verfügbar:
PURL: http://resolver.sub.uni-goettingen.de/purl/?dariah-2013-2
URN: http://nbn-resolving.de/urn.nbn.de.gbv:7-dariah-2013-2-1

Bibliographische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Daten sind im Internet über <http://dnb.ddb.de> abrufbar.

Erschienen in der Reihe
DARIAH-DE Working papers

ISSN: 2198-4670

Herausgeber der Reihe
DARIAH-DE, Niedersächsische Staats- und Universitätsbibliothek

Mirjam Blümm, Stefan Schmunk und Christof Schöch

Abstract: The workshop on Quantitative Text Analysis for Literary History was the first in a series of DARIAH-DE expert workshops and took place from November 22 to 23 at the University of Würzburg, Germany. It brought together experts in the computational analysis of collections of literary texts from France, Germany, Poland and the US. This report provides some context on the DARIAH expert workshops and introduces the specific goals of the workshop reported on here. Then, it summarizes the major issues raised by the participants in their initial statements and debated in the ensuing discussions. Finally, it describes the key results from the workshop, which include advances in the areas of data, tools and methods for quantitative text analysis.

Keywords: Digital Humanities, Literaturgeschichte, Textanalyse

Quantitative Text Analysis for Literary History – Report on a DARIAH-DE Expert Workshop

Christof Schöch, Fotis Jannidis (Univ. of Würzburg)

Abstract

The workshop on *Quantitative Text Analysis for Literary History* was the first in a series of DARIAH-DE expert workshops and took place from November 22 to 23 at the University of Würzburg, Germany. It brought together experts in the computational analysis of collections of literary texts from France, Germany, Poland and the US. This report provides some context on the DARIAH expert workshops and introduces the specific goals of the workshop reported on here. Then, it summarizes the major issues raised by the participants in their initial statements and debated in the ensuing discussions. Finally, it describes the key results from the workshop, which include advances in the areas of data, tools and methods for quantitative text analysis.

DARIAH-DE
Working Papers

Christof Schöch & Fotis Jannidis (2013): "Quantitative Text Analysis for Literary History – Report on a DARIAH-DE Expert Workshop", *DARIAH-DE Working Papers* Nr. 2. Göttingen: DARIAH-DE.
URN: urn:nbn:de:gbv:7-dariah-2013-2-1.

Die *DARIAH-DE Working Papers* werden von Mirjam Blümm,
Stefan Schmunk und Christof Schöch herausgegeben.

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Introduction

The workshop on *Quantitative Text Analysis for Literary History* was the first in a series of DARIAH-DE expert workshops and took place from November 22 to 23 at the University of Würzburg, Germany.[1] It was organized by Fotis Jannidis and Christof Schöch of the Department for Literary Computing, University of Würzburg, and brought together experts in the computational analysis of collections of literary texts from France, Germany, Poland and the US.

This report first briefly provides some context on the DARIAH expert workshops and introduces the specific goals of the workshop reported on here. Then, it summarizes the major issues raised by the participants in their initial statements and debated in the ensuing discussions. Finally, it describes the key results from the workshop, which include advances in the areas of data, tools and methods for quantitative text analysis.

Pushing the Edge – Expert Workshops in DARIAH-DE

Expert workshops in DARIAH-DE focus on the critical analysis and further development of methods and tools relevant to a relatively specific issue in digital humanities research, with the aim of making some concrete progress in this area. The issues addressed may or may not be directly connected to DARIAH services or efforts. The participants primarily targeted are confirmed researchers with a solid record in humanities and/or digital humanities research as well as experience with using, enhancing and developing methods, tools, or datasets. The goal is to bring a small group of such experts together, let them exchange their vision of the current state and most relevant issues of the field in question, and offer them the opportunity to collaboratively push their own research on digital methods, tools, or datasets one step further. Expert workshops generally combine a limited amount of informal project presentations from the participants themselves with large room for discussion and work in small teams on specific methods, tools and/or datasets.

Text Analysis for Literary History – Aims of the Expert Workshop

The expert workshop on *Quantitative Text Analysis for Literary History* focused on methods and tools in the domain of quantitative analysis of large text collections in the context of literary history. The aim of this workshop was to consider recent developments and issues in quantitative text analysis and their relevance to the way we understand and write literary history. Such developments include the refinement of authorship attribution studies, of clustering techniques for literary genre analysis, of topic modeling, intertextual analysis, and of other stylometric or computational approaches to literary texts.

[1] For more information about DARIAH-DE events and activities, see http://de.dariah.eu/events (in German).

One of the issues is on what level of analysis computational approaches to text analysis can usefully focus, i.e. surface features of texts, such as type or token frequencies, or higher-level semantic, morphological or structural features, and how to mediate between these two poles. Another issue is how to make use of text which either contains structural encoding, is enriched with linguistic annotations, or provides bibliographical metadata to be taken into account in the analysis. A third issue concerns the way evolution over time and across genres can be traced or modeled using computational approaches and very large collections of text.

The workshop focused on the relevance of this type of issues for the analysis of literary or other genres, the long-term evolution of stylistic or narrative features, and related topics, with a focus on their bearing on literary history. With these aims in mind, this workshop brought together a small group of scholars from literary studies and information sciences who are currently employing computational approaches to the analysis of large collections of literary texts and who are interested in further developing the necessary methods, tools, scripts and algorithms for such approaches.

The workshop aimed to consider recent issues in quantitative text analysis and their relevance to the way we understand and write literary history.

I. Who's doing what – Initial statement session

The larger part of the first day was devoted to initial statements by each participant on their current work in quantitative text analysis, on the challenges and opportunities they currently see for the field, and the issues they would like to work on during the remainder of the workshop.

Serge Heiden (ENS Lyon, France)

As part of the Textométrie team, Serge Heiden has been involved in the development of the open-source quantitative text analysis platform TXM.[2] He presented the main characteristics of TXM: the platform combines quantitative methods (frequency lists, factorial analysis, classification, collocates, etc.) with qualitative ones (text browsing, full text search, kwic concordances, etc.); it uses efficient open-source technologies to implement these methods, R for quantitative analyses and CQP for qualitative / linguistic queries.[3] It uses standard software frameworks, such as Java / OSGi / the Eclipse RCP framework and processes standard textual formats, such as Unicode, XML, TEI. It provides scripting facilities (in Groovy) to automate sessions and permit the user to adapt import scripts, for TEI import for

[2] TXM, version 0.6, 2012, http://textometrie.ens-lyon.fr/?lang=en.

[3] See http://cran.r-project.org/ for the R framework and http://cwb.sourceforge.net/ for CQP.

example. Finally, it is deployed as a free desktop application or web portal and is made available under an open-source licence.

The discussion revolved around the question of how the strengths of TXM – which lie in the areas of corpus management and linguistic annotation as well as in the fact that TXM already permits running R scripts from within the program – could be combined with the capabilities of Eder & Rybicki's "stylo" script.[4] It was decided that building this integration, which would allow researchers to do rich and detailed searches over linguistically annotated corpora and use the resulting feature lists as input for the stylo script, should be one of the activities of the working phases of the meeting. Possible collaboration between the TXM project and TextGrid was also discussed, considering that the two applications share the Eclipse framework as their basis and are in several ways complementary in functionality.

Fotis Jannidis (Würzburg University, Germany)

In his statement, Fotis Jannidis put particular emphasis on the fact that in many stylometric investigations, not only an author signal, but also a very strong genre signal is found. This means that stylometry is a methodology promising an entirely new means, on a very large basis of materials, of establishing groups of texts similar by genre, and in this way to rewrite literary history. Fotis Jannidis' and Gerhard Lauer's findings in stylometric studies of German Literature between 1750 and 1900 raise the the question of how, on the basis of authorship attribution techniques, larger trends relevant to literary history could be identified, something for which factors like genre and gender seem to be highly relevant: how do larger generic clusters shift over time, how do new generic clusters emerge, how does a broader view of the literary production challenge the labels and divisions we use to classify genres (such as Bildungsroman or gothic novel) and literary eras (such as Enlightenment or Romanticism).

One methodological issue discussed in this context was whether wordlists from topic modeling, which would possibly capture the common thematic ground of texts belonging to a given genre, could be used as the basis for stylometric analyses using Eder & Rybicki's scripts. The results of such topic-based stylometric classifications could then be compared to more traditional stylometric classifications based on most frequent words.

Maciej Eder (Pedagogical University of Kraków, Poland)

In his statement, Maciej Eder pointed to the fact that several recently introduced stylometric methods (such as those introduced or used by Burrows, i.e. Principal Components Analysis, Cluster Analysis, and Delta) are very intuitive and easily-applicable to literary studies. However, these methods have important limitations, above all a lack of a built-in measure of

[4] See below, and http://sites.google.com/site/computationalstylistics/, for details on the "stylo" script/package for R.

validity or reliability of the obtained results. More sophisticated machine-learning algorithms suitable for classification tasks and derived from the sciences, (such as Support Vector Machines, Decision Trees, or Penalized Regressions), are usually ignored in literary-oriented case studies, because understanding them requires sound statistical knowledge. Bridging this gap by combining advanced methods with suitable visualisations is one of the key challenges for computational text analysis in the years to come.

The discussion revolved around the question of how we can deal with the "black-box" problem. Computational text analysis should rely on increasingly sophisticated statistical methods, but ways need to be found to make the underlying process transparent to humanities researchers, who need to be able to understand, modify, and trust the procedures they use. The bootstrap consensus tree, which the stylo script includes, is one solution to this issue, creating a cross-validation of many cluster analysis runs with a suitable visualisation; however, some information on the exact distances of the individual texts to each other is lost in the process; one possible solution to this may be to add the bootstrap/validation information to the more traditional dendrogram visualisations.

Gerhard Lauer (GCDH, University of Göttingen, Germany)

Gerhard Lauer reported on recent work on corpora of German literature, with a focus on drama and narrative around 1800. One particularly interesting case are the writings of Heinrich von Kleist, because they present a challenge to traditional attempts of literary history classification: Kleist is neither a true Romantic, nor a true Enlightenment author, it seems. While stylometric classification shows Kleist closer to other Enlightenment authors than to Romantic authors, for specific plays this is not the case, with some variation in the results depending on the settings used, which again raised the question of the validation of results also brought up by Maciej Eder.

During the discussion, and in the context of validation, the issue of a comparative corpus for stylometry emerged. Such a corpus, which should be multilingual and on which different researchers could work using various methodologies, would facilitate the cross-method validation of results and foster collaboration and comparison of the success of methods between various groups of researchers in stylometry. Gerhard Lauer argued that this could also help move stylometry into the realm of comparative literary history, with research into mechanisms and trends of innovation and canonization as well as the stability or fluidity of genre boundaries.

Allen Riddell (Duke University, USA)

Allen Riddell reported on research in the area of quantitative analysis of the British Novel between 1800 and 1836, based on a collection of nearly 3000 novels. Taking Moretti's chapter on genre clusters as his starting point, Riddell showed that topic modeling can be a useful way of investigating the question of novelistic (sub)genre (such as gothic, national tale, Bildungsroman). Received accounts of novelistic genre seem inadequate once topic modeling is used to perform unsupervised grouping of novels based on their topical affinity.

One of Riddell's main points, which was hotly debated during the discussion, was a tale of caution regarding the way corpuses are usually built for stylometric or topic modeling studies. He insisted on the necessity of creating randomized samples in order to establish statistically valid results. This, in turn, is only possible with solid bibliographic information on the entire novelistic production of any given period, based on which random sampling could be done. Then, all randomly selected novels need to be made available in full-text digital form. Because random sampling means only a small part (around 5-10%) of the actual novelistic production needs to be digitized, this is also a pragmatic strategy in terms of digitization efforts which, however, are rarely conducted with the aim of establishing a random sampling, but are based on arbitrary and/or pragmatic criteria of the digitizing institutions. The discussion then revolved mostly around the possible paths and difficulties in establishing such a randomized corpus for several European languages in a way that would allow comparative studies of literary genres and a benchmark corpus for testing and comparing stylometric tools and algorithms.

Jan Rybicki (Jagellonian University, Krakow, Poland)

Jan Rybicki started his statement quoting Wincenty Lutosławski who, in 1898, probably used the term stylometry for the first time in its modern sense, defining it as "the measure of stylistic affinities". This led Rybicki to raise the issue of what kind of affinities these are, and whether they are related, apart from literary style, to gender, chronology, tone, or even quality. This also raised the issue of the relation between stylometry (which primarily classifies styles) and stylistics (which describes styles), a relation which is more distant than the term stylometry suggests. Both domains of inquiry, Rybicky said, should use each other's findings and methods more readily than they currently do.

Connecting stylometry and stylistics would also be a way of connecting 'traditional' humanistic inquiry with an approach germane to the digital humanities. In addition, as emerged frequently later on as well, it is urgent to take the special nature of natural language into account when devising statistical classification methods, which are today frequently based on algorithms or procedures borrowed from disciplines not dealing with language.

Christof Schöch (University of Würzburg, Germany)

Christof Schöch presented his current stylometric work, concerned with investigations into parameter setting and its impact on classification of such texts by author, genre, or form, using Eder & Rybicki's stylometric scripts for R, specifically in French classical drama, and around the Molière-Corneille controversy. One issue standing in the way of reliable results concerns the degree to which such authorship classification tasks are influenced by genre (here, comedy or tragedy) and form (here, verse or prose).

In this context, Christof Schöch raised the issue of methodological improvements to the disentanglement of author, genre and form signals from the data, and that of how similar investigations into different collections (different languages, types of text, etc.) could be coordinated and made comparable, so that a useful body of knowledge and experience is created. One of the questions discussed in this context was what DARIAH had a role to play in such a coordination of efforts, on the one hand, and what could be accomplished via a mailing-list or wiki centered around the community of stylometry researchers.

II. Challenges identified, and solutions developed – the working phase

Based on the initial statements by the participants and the discussions of each of them, the participants identified several challenges the field of quantitative text analysis for literary history currently faces. The three most important of them defined the focus of the work done during the end of the first day and during the second day of the workshop. These issues where the following:

1. On the level of data, conceptualize a "European Comparative Corpus" of literary texts, in order for reliable, valid and comparable stylometric results for large-scale trends and for benchmarking of methods to be possible.
2. On the level of methods, investigate the use of topic models in stylometry, with a view to enhance the scope of possible types of features to be used in stylometric analyses.
3. On the level of software and methods, integrate the "stylo" script into the TXM environment, to make it possible to use a wide range of linguistic features as the basis of stylometric analyses.
4. In addition, think about some smaller enhancements to the "stylo" script by Eder & Rybicki.

These goals and issues guided the work during the remainder of the workshop. The next section outlines results from the work during the workshop.

European Comparative Corpus for Stylometry – A Proposal

In order for stylometric studies to produce reliable and valid results concering large-scale trends in literary history, the establishment of a large reference collection of literary texts is necessary.

The time period which currently seems best suited to start such an endeavor is that between ca. 1780 and 1850, for the following reasons: the level of availability of good-quality digital facsimiles or full-text versions is relatively high, the bibliographic research into the total literary production of the period is relatively advanced (with variation depending on language, however), and the total production has not yet risen to unmanageable numbers. Such a corpus should be constructed in the following way, each step being documented in detail:

1. Selection of a 5-10% sample of texts based on this record, in 10-year slices, excluding texts with a length of less than 5000-10000 words. Assuming a per-decade production of about 1000-2000 texts, this would mean 50-200 texts per decade. Random selection of these texts seems to be beyond what is currently possible, especially with a view to a lack of complete bibliographic records which would be a precondition for random selection.[5][4] The corpus could be built first for novels only, and later expanded to include other literary genres.
2. Quality assurance for existing full texts, and digitization and double-keying of those texts where searchable full texts are not yet available. The latter is expected to be the case for an estimated 60% of the texts, for English; probably more for German, French, Polish and other languages.
3. Collection and storage of detailed metadata about each texts, particularly with respect to descriptive metadata (author, date of writing/publication, genre, sub-genre, literary epoch, gender of author, translation or not), technical metadata (text length in words, part of the randomized sample or not, edition used in establishing the text), and administrative metadata (OCLC number, URI/DOI, in-corpus ID), using a simple metadata scheme such as Dublin Core (if sufficient) or a more specialized schema such as OLAC.
4. Linguistic annotation of the corpus: orthographic normalization, tokenization, lemmatization, part-of-speech tagging. These procedures should be done based on tools or services which have been trained with appropriately selected materials from the same time period.
5. Storage of all texts and metadata in a subversion repository (SVN) or on Github, in order for continual or periodic improvements of the corpus to be manageable.

For the following issues, a coherent strategy needs to be established, and the necessary expertise needs to be established:

1. The amount and nature of normalization of the texts;

[5] The availability of bibliographies of literary production varies for different national literatures, and needs to be investigated more thoroughly.

2. The way translations and multiple editions are dealt with (include or exclude, use available editions from a time after initial publication, etc.);
3. The way in which the corpora for different national literatures are being made comparable in size, generic scope, or other features. (This would be solved by the random selection procedure, which makes the corpora representative).
4. How can the quality of the texts (OCR, double-keying), of orthographic normalization, and of linguistic annotation (tokenization, lemmatization, POS-tagging) be assessed and guaranteed?

The availability of such a corpus is a basic and essential requirement for many different individual projects, and is in fact an infrastructural issue. Therefore, a consortium such as DARIAH, which builds the infrastructure components for research in the humanities and which has a strong competency in the domain of research data, should be an active force in the establishment of such a corpus. DARIAH could help coordinate the efforts, help with identifying funding opportunities in different countries, and make recommendations for the metadata scheme used and the data storage strategy pursued. Also, DARIAH could provide server space or virtual machines for processing of texts. Finally, DARIAH could help the digitization efforts by bringing together its many partner institutions with library collections and digitization infrastructures.

A task force should be established to identify relevant and interested scholars from around six to eight European countries, and national funding opportunities should be used for the digitization of each national language sub-corpus.

Integration of the stylo script into the TXM environment

A second activity during the workshop, driven for the largest part by Serge Heiden and Maciej Eder, was to integrate the stylo script into the TXM environment. TXM provides capabilities for detailed multi-level linguistic annotation and querying, while the stylo script allows for stylometric analyses using various distance measures and visualisations. Combining the two tools allows researchers to do rich and detailed searches over linguistically annotated corpora and use the resulting feature lists as input for the stylo script.

This integration was already functioning by the end of the workshop. It allows the researcher to import and preprocess TEI-files from TextGrid to create a TXM corpus, to make complex CQP queries and use the resulting list of features (for example, all articles and prepositions, or all instances of auxiliary verbs, etc.) as the wordlist to be used by the stylo script. In order to do this, the feature list is sent to R, where the stylo script is run to build the frequency and distance tables, using the feature list instead of the standard wordlist, and the resulting graph is displayed inside TXM. This integration is an extraordinary expansion to the stylometric toolbox, significantly enlarging the types of linguistic features available for stylometric distance calculations beyond words, letters, and word/letter n-grams. A "stylo-for-txm" variant of the "stylo" script will be released in due time.

Combining Topic Modeling with stylometric analysis

The aim of this activity was to facilitate the combination of topic modeling with stylometric distance measuring techniques. To this end, participants devised a workflow in which topic modeling over a large collection of literary texts is done with MALLET.[6] Settings need to be adapted appropriately so that useful material is generated; notably, this means enlarging the number of terms per topic beyond the usual cut-off line of twenty or thirty terms to at least 100 terms per topic. A table of relative frequencies is then constructed in which instead of texts and words, texts and topics form the two axes, and instead of normalized word frequencies, the topic's weights are used for the clustering or nearest neighbour calculations. This table is transformed into a format that the "stylo" script can use and the usual analyses are then performed on it.

One theoretical advantage of the topic-based clustering method over the Zeta method, which is equally based on a predominantly semantic differentiation of texts, is that while the Zeta method only contrasts preferred and avoided words for one set of texts as opposed to some reference corpus, the topic-based approach yields scores for each topic and each text. Preliminary results from several test runs show that with this method, the author signal stays very strong, so that for a set of British novels from the eighteenth and nineteenth century, for example, the grouping happens in a way quite similar to more traditional stylometric methods. However, more subtle differences exist, and need to be assessed more thoroughly in the future.

The benefit of this combination of topic models with stylometry, just as is the case for the integration of linguistic features into the stylometric calculations, is that with more and more features being used as the basis for stylometric calculations, results from different methods and based on different types of evidence can be compared, and in this way the reliability of various methods can be assessed.

Other outcomes

During the course of the workshop, several ideas for further development of Eder & Rybicki's stylometric scripts for R were discussed. Some were implemented right away, and others have already been implemented at the time of this writing. First, the distance measure table, which is calculated from the word frequency table, and used for the creation of the dendrograms, can now optionally be saved as a separate file; in this way, this data becomes available for further quantitative analysis. Second, the text files to be used in a given

[6] See http://mallet.cs.umass.edu/.

stylometric run can now be selected from the corpus folder based on a list of files specified in a separate file; such files can easily be created on the basis of a user-selection, which makes it much easier to create similar analyses with varying subsets of a given text collections. Finally, it is now possible to color the dendrograms based on the author label, in a way similar to that of the coloring of the bootstrap consensus trees; this makes the visual identification of patterns in the results easier. These enhancements of the stylo script have been incorporated into version 0.4.8 of the script.

Conclusions

To everyone involved, the success of the event has been quite obvious, and we hope that the present report also shows this. All participants have brought their specific expertise to bear on the issues at hand, allowing the event to produce considerable advances on the methodological as well as technological level. The enlargement of the types of data that become available to stylometric classification tasks is probably the most important one, and is both a methodological progress (the implications of this new possibilities have been reflected upon and tested) and a technological progress (applications have been developed or data brought together in order to concretely enable these innovations.

In addition, the conceptualization of a "European Comparative Corpus for Stylometry" will hopefully foster useful activities in the domain of corpus construction. At the same time, it has become very clear to everyone that a lot of effort will still be needed to help the field mature further: more and better text collections and ever more flexible and more usable tools will go hand in hand with more widespread uptake of and experiences with computational methods in the analysis of literary texts.

As far as the instrument of the "expert workshop" itself is concerned, the event has confirmed that small meetings of dedicated practitioners, some with a focus on humanistic enquiry, some with a strong background in tool development (and, in this case, statistics), produces new ideas and allows for these ideas to be realized and tested right away. We are looking forward to the next DARIAH-DE expert workshops which will be organized soon for other domains in the digital humanities.