Scientific report

Second and third workshop of the NeDiMAH working group Using Large-Scale Text Collections for research

Workshop Tuesday April 1st, 2014: **'Corpora'**

Workshop Wednesday April 2nd, 2014: **'Research'**

1) Summary

The workshop "Corpora" dealt with the interface between linguistic annotation and textual annotation for historical and literary etc. research. It especially dealt with the interface between linguistic annotation and textual annotation for historical and literary etc. research. It aimed to bring together corpus builders and those corpus users other than linguists. Corpus builders were able to inform the participants about such things as the kind of requests for functionality they get from non-linguists and their answers and advice to those scholars. Non-linguistic corpus users were invited to talk about the different ways in which they have gathered their own corpus, whether they especially built a private corpus for their use, and why the have taken such a step, or what they would like to see in a corpus before they will actively start making use of it.
We invited informative short talks in order to generate an open and exploratory discussion between participants with different disciplines as a background. Discussions arose on such topics as samples versus complete texts, standard functionality for text research, the need of export options to get the texts to the scholar to be used in his or her stand-alone tools, in opposition to the adding of tools to a corpus, strategies for and issues in building a multilingual reference corpus for text analysis, and much more.

The workshop "Research" focused on new kinds of analysis of large text corpora explicitly from the perspective of literary or historical research questions. In this context, for example, the issue of validation of results gains importance, not primarily in the sense of statistical measures of validity or robustness, but rather in the sense of interpretation and validation of results in relation to literary or historical etc. knowledge.
We invited short papers describing a concrete case in which the use of (relatively) large-scale text collections has lead to new insights that could not have been gotten using non-digital methods. These kinds of case studies have a high impact on the willingness of humanities scholars to broaden their toolkit to digital and computational methods.

2) Description of the scientific content of and discussions at the event

After the explanation of the aims of the meetings by the chair, Karina van Dalen-Oskam, several attendants presented themselves and their main project in short presentations: Martin Wynne about the Oxford text Archive and the British National Corpus, Tom van Nuenen about his PhD-research in to modern pilgrim texts, Susan Schreibman on the Irish project Letters of 1916, Puck Wildschut about her PhD-research into Nabokov's style, (and on the second day:) Florentina Armaselu about the Luxembourg Centre Virtuel de la Connaissance de l'Éurope, and Karina van Dalen-Oskam about the project The Riddle of Literary Quality.

The first long presentation was given by **Christian Thomas**. He spoke about 'The CLARIN-D Service Centre at the BBAW: Corpora, Tools, Methods, and Best Practices for text-based interdisciplinary research'. This CLARIN-D Service Centre (clarin.bbaw.de) ensures the availability and long-term preservation of German historical and contemporary text corpora, and lexical resources provided by the Zentrum Sprache (Language Centre) at the BBAW. CLARIN-D (www.clarin-d.de) is the German section of the CLARIN European research infrastructure federation, and develops a web and centres-based research infrastructure, building on the expertise of currently nine service centres in major research institutions. At the BBAW, historical and contemporary text corpora, as well as lexical resources are compiled and integrated into the CLARIN-D infrastructure, where tools for further analysis of the data are provided and their long-term preservation is taken care of. A federated content search and sophisticated retrieval facilities allow for cross-collection queries in the greater context of all corpora available in CLARIN-D.

The center hosts several corpora. First, The Deutsches Textarchiv (DTA, www.deutschestextarchiv.de), funded by the German Research Foundation (DFG), provides a broad selection of more than 1300 significant German works of various disciplines. Their time of origin ranges from ca. 1600 to 1900. In addition, the DTA core corpus is extended by currently 473 high-quality textual resources provided by cooperating projects or curated from existing text collections such as Wikisource and Project Gutenberg. The lexical information system of the Digital Dictionary of the German Language („Digitales Wörterbuch der deutschen Sprache", DWDS) is accessible to all users via the internet (www.dwds.de). Besides several lexical resources such as the DWDS-Wörterbuch, the Wörterbuch der deutschen Gegenwartssprache (WDG), or the Deutsches Wörterbuch by Jacob and Wilhelm Grimm, it encompasses a large, linguistically anntotated German corpus of 20th and 21st century texts containing about 1.8 billion tokens.

Jan Rybicki's presentation went into 'The Benchmark Corpus of English Literature'. He sketched the background of the need for a benchmark corpus: a corpus that can be used and reused with the specific purpose to check how well new tools perform. However, many other wishes present themselves to the builder of such a benchmark corpus. Its present early version has been made, as had been previously agreed with Corpus Co-makers, with the following requirements for the entries (problems encountered are listed alongside each requirement): (1) "Representativeness" of the selection of texts; this is obviously very vague and the inclusion, or not, of individual authors and texts will need to be discussed case-by-case; (2) Variety of sub-genres and high/lowbrow proportions within the general genres of 19th-early 20th narrative prose; the problem to be solved here is the extent of the variety. For instance, should children's literature be included? How to balance the "good" and "bad" literatures? (3) The entries must be in the public domain; this is usually ascertained by adding 75 years to the author's date of death. In some cases, texts have been placed in

public-domain collections before this requirement was fulfilled (e.g. Virginia Woolf at gutenberg.org). (4) Three texts per author. This limits the eligible authors (those famous and significant due to a single book, for instance) and creates the additional problem in cases of *embarrass de richesse* – which three books to choose from the output of very active novelists like Dickens. A working rule was observed to select an early, a middle and a late novel. (5) Adequate representations of the two genders. In the first selection presented here, as many as one third of the authors are female. This might be a problem with the final requirement, namely that (6) The English corpus be as "similar" as possible to future benchmark corpora in other languages (for instance, the gender imbalance might be more marked in other cultures). (7) That the corpus be not too "easy" from a stylometric point of view. This selection's attributive success (using the *classify* function in *stylo*) is currently around 80%.

René van Stipriaan went into the large Dutch project Nederlab, in which a large diachronic corpus of Dutch is being built. He sketched the background of the project, taking account of issues such as representativeness and difficulties arount digitizing older text material. From the user perspective, subcorpus selection and other issues were dealt with.

Allen Riddell's European Novels Research Corpus has a totally different approach. The project aims to collect all surviving novels published between 1770 and 1901. This period witnessed the publication of as many as 50,000 novels in the British Isles alone. The first phase of the project will collect a random sample of 100 novels drawn from the 2,903 novels published between 1800 and 1836 in the British Isles (as documented by Garside and Schöwerling (2000)). A second corpus focuses on novelistic genres and gathers together a large but non-random sample of novels classified as "gothic", "silver fork", and "national tale" (all genres appearing in the period 1800–1836). Digitizations of novels are gathered from a variety of sources, principally from the Internet Archive, Google Books, and the Corvey Collection. As digitizations of novels from the Corvey Collection are often poor, a significant number of these novels have been transcribed by human readers. The transcription is facilitated by Amazon Mechanical Turk at a cost per page of less than €0.20. Datasets will be released with fixed versions to enable reproducible research but still permit the correction of transcription and OCR errors over time. (The Supreme Court Database, which concerns the US Supreme Court, is an inspiration in this regard.) All material associated with the corpus will be available for bulk download. An early version of the interface will shortly be available at http://novels.abr.webfactional.com. The project is unorthodox in that it abandons the ambition of error-free transcription or pristine TEI encoding in favor of building upon mass digitization projects to assemble a large, reasonably accurate, representative sample of cultural production.

Gordan Ravančić (through videoconferencing) explained his work into a Digital quantitative approach in investigation of daily life of medieval city using examples from Dubrovnik. The core of his talk was to show what valuable quantitative information can be extracted from serial archival sources. In further qualitative interpretation such quantitative data can lead to a new comprehension of pre-modern reality. Namely, with a methodology of creating a digital database(s), i.e. structuring serial archival sources into (relational) database, one can grasp much more information about life patterns of our predecessors than it is possible with usual historiographical qualitative and descriptive methods. His first case was an investigation of the preserved testaments from which one can easily get insight about contemporary economic, social and spiritual relations among inhabitants of medieval Dubrovnik. It may be

said that testaments as private legal documents are among the best sources for study of economic, social, legal, cultural and spiritual life in medieval Dalmatian communes. The analysis focuses on a relatively small sample of 432 testaments, of the total of over 900, from this period. The examined testaments are from the following years: 1295, 1296, 1325, 1326 and 1348. While it is obvious that the sample is relatively small when compared with the total number of testaments kept in the Dubrovnik archives, this study should reveal certain changes in the distribution of bequests over the examined time period. These results should not be regarded as 'absolute' but they do reflect certain social, economic and organizational trends in the contemporary Dubrovnik communal social system, as well as concerning their spiritual mentality. His second case was a research about crime patterns in late medieval Dubrovnik. It would be interesting to investigate to what extent crime frequency followed pre-modern rhythms of labor, leisure and calendars of festivities, at what locations most crimes occur and the social position of perpetrators and victims. A sample of 250 recorded criminal cases has showed that, undoubtedly, pattens can be established between certain crimes and places of their occurrence, as well as the fact thas there was a notable congruence between certain crimes and some social groups. All these valuable data cannot be revealed without a quantitative approach in historical research, and digital methods can speed up and help on our way towards more comprehensive reconstruction of daily life in a pre-modern medieval city.

Maciej Eder not only presented a paper about 'stylometry, big data, and possible bottlenecks', but also gave an impromptu hands-on workshop using the Stylo package in R. He states that "Big Data" means access to previously unheard-of amounts of data; at the same time, however, this presents non-trivial challenges. One of the most obvious (and the most painful) is that one cannot reliably verify the quality of hundreds of texts to be analyzed. Even worse: in a large collection of files, one cannot simply check if these files really contain textual data. Stylometry, or assessing stylistic similarities between literary texts using statistical methods, is quite sensitive to the problem of untidily prepared corpora. Very short text samples, numerous misspelled characters, different orthographic conventions -- thes are some of the the factors that affect the results substantially. The bigger a corpus, the more chances there are to deal with these issues. To address this problem, Eder chose one of the best curated collections of texts for a test, the corpus of Ancient literature provided by the "Perseus Project" database (http://www.perseus.tufts.edu/hopper/opensource/download). It is small enough (1127 texts) to be controllable and/or manually emendated, and at the same time big enough to serve as a simulation of "Big Data". The results of a stylometric experiment using network analysis techniques show that there is no easy way to analyze this collection algorithmically. E.g. a trivial problem of separating texts in Latin, Greek and English becomes challenging when a corpus becomes large.

Emma Clarke gave a very clear demonstration of the pro's and cons of 'Topic Modelling Transactions of the Royal Irish Academy 1800-1899'. She presented the pipeline which was built on that corpus of quite noisy text, which contains articles with a high degree of specialisation; and discusses the effect of different attempts to regularise and normalise the machine mediation of the insight. The goal of the research was to get a more coherent view of the word associations that arise from the semantic, rather than the positional in the document. Her main conclusions were that (1) The cleaning process has a direct impact on the topics and their stability. Namely, the choices that are made in the data cleaning process

can directly impact upon the topic output; (2) That transferring parallel cleaning rules can be hazardous. Taking the earlier methodology used by Goldstone and / or Underwood and implementing it 'out of the box' was not a viable option. She made a number of interesting observations showing how topics shift over the nineteenth century. She drew conclusions by taking topics and graphing their presence within the corpus, while observing a timeline of cultural and historical events over the nineteenth century time-frame. All in all, with all it's difficulties, topic modelling as a form of distant reading is an innovative and exciting method that could be best employed as a discovery tool for further research on corpora such as Royal Irish Academy journals.

After this, **Christof Schöch** and **Fotis Jannidis** sketched plans about a COST Action titled 'Computational Stylistics for European Literary History' and about the research questions related to this and to the benchmark corpus that Rybicki presented earlier. This led to a general discussion on all the topics presented.

3) Assessment of the results and impact of the event on the future directions of the field

The meetings had several important results. The first presentation (Thomas) gave a thorough insight into what building, sustaining and serving a large corpus means from the perspective of the makers. It also revealed how important the builders find the usage that scholars would want to make of their corpus, and how difficult it is to get the information needed.

Rybicki's presentation on the benchmark corpus led to intensive discussions and new insights into how to limit the number of variables: the testing aim of the corpus should be primary, and any additional wishes (genre differentiation, gender balance, etc.) should be of secondary importance to make sure these do not reduce the 'benchmarking' possibilities in any way.

Riddell's presentation on his novel corpus was insightful in several respects, as became clear in the discussion. The idea of representativeness, which is nowadays seen as almost impossible to reach in a large, general corpus, in fact seems realisable in the way Riddell shows it. Another important result from his presentation was the way statistics can be used to evaluate the corpus as being representative, making use of bibliographic and other kinds of data available.

Ravančić's presentation mostly led to questions from a much broader perspective than the one he sketched here: do we have similar information from other cities, could they be easily compared with the Dubrovnik ones, etc. The historical and social perspective of his talk added a new perspective to the earlier linguistic and literary ones in the workshops.

Eder's discussion of big data, stylometry, and the demonstration of R and the Stylo package, led to new insights for many participants. For the linguistic corpus builders, the fact that stylometrist only use raw text files, was kind of shocking. In fact, neat TEI-files are a hindrance for stylometrists - they would first want to remove all the tagging before doing their analysis, using their own appropriate tools. Eder also made the participants aware of the many ways in which multivariate analysis can go wrong when the underlying data and/or the workings of the tools are not very clear to the users. This sketches the need for many checks of every step in the process, showing indeed how difficult it is to estimate the reliability of measurements done on big data of which the form and quality is not clear.

Clarke's main point in fact proved to be related to expectations about what digital tools can bring to humanities scholars and how they may be unrealistic, expecting too much. But when we downscale our expectations, the new tools can indeed help the researcher to find and see much more than (s)he cold ever before and thus get new ideas, new perspectives, for next steps in the research.

The final discussion brought all the topics dealt with together. Participants listed the issues that were new for them, and the group discussed in what ways the insight could be disseminated further: a volume of articles, or a kind of manual addressing a selection of topics such as: how to build a corpus for specific research? How to use existing corpora for specific research? What kind of tools exist for e.g. topic recognition, style analysis, etc.? The organisers of the workshop are currently building on the suggestions that were gathered and hope to come up with a suitable plan in this direction soon.

4) Final programme of the meeting

Workshop Tuesday April 1st, 2014: **'Corpora'**
Workshop Wednesday April 2nd, 2014: **'Research'**

Second and third workshop of the NeDiMAH working group Using Large-Scale Text Collections for research.
http://drupal.p164224.webspaceconfig.de/workgroups/using-large-scale-text-collections-research

Location: Universität Würzburg, Am Hubland, Bau 8, room "ÜR 20".
Organizers: Karina van Dalen-Oskam, Fotis Jannidis, Christof Schöch
Contact: karina.van.dalen@huygens.knaw.nl

Programme Tuesday April 1st, 2014: **'Corpora'**

9.30 – 9.35:
Karina van Dalen-Oskam (Huygens ING / University of Amsterdam), **Introduction: aims and tasks**

9.35 – 10.15:
Round of introductions including some **5-minutes' presentations**

10.15 – 10.45:
Frank Wiegand, Christian Thomas: **The CLARIN-D Service Centre at the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW): Corpora, Tools, Methods, and Best Practices for text-based interdisciplinary research**

10.45 – 11.15: Coffee break

11.15 – 11.45:
Jan Rybicki: **The Benchmark Corpus of English Literature and Computational Stylistics for European Literary History**

11.45 – 12.15:
René van Stipriaan: **Nederlab, building a large diachronic corpus of Dutch**

12.30 – 13.30: Lunch

13.30 – 15.00:
Allen Riddell: **Building a European Novels Research Corpus**

15.00 – 15.30: Tea break

15.30 – 17.30:
Discussion

Programme Wednesday April 2nd, 2014: **'Research'**

9.30 – 9.35:

Karina van Dalen-Oskam (Huygens ING / University of Amsterdam), **Introduction: aims and tasks**

9.35 – 10.15:

Round of introductions including some **5-minutes' presentations**

10.15 – 10.45:

Gordan Ravančić: **Digital quantitative approach in investigation of daily life of a medieval city - examples from Dubrovnik**

10.45 – 11.15: Coffee break

11.15 – 11.45:

Maciej Eder: **Stylometry, big data, and possible bottlenecks**

11.45 – 12.15:

Emma Clarke: **Topic Modelling Transactions of the Royal Irish Academy 1800 – 1899**

12.30 – 13.30: Lunch

13.30 – 15.00: The planned COST Action Computational Stylistics for European Literary History

-Christof Schöch: **General presentation of the planned COST action**

-Fotis Jannidis: **Research questions enabled by the COST Action benchmark corpus**

-Discussion.

15.00 – 15.30: Tea break

15.30 – 17.30:

General discussion, planning of next steps in the NeDiMAH working group

5) List of participants

Florentina Armaselu, Centre Virtuel de la Connaissance de l'Europe, Luxemburg

Emma Clarke, National University of Ireland, Maynooth, Ireland

Maciej Eder, Pedagogical University, Krakow, Poland

Fotis Jannidis, University of Würzburg, Germany

Gordan Ravancic, Croation Institute of History, Croatia (via videoconferencing)

Allen Riddell, Dartmouth College, USA

Jan Rybicki, Jagiellonian University, Krakow, Poland

Christof Schöch, University of Würzburg, Germany

Susan Schreibman, National University of Ireland, Maynooth, Ireland

Christian Thomas, BBAW, Berlin, Germany

Karina van Dalen-Oskam, Huyghens Institute, The Hague, The Netherlands

Tom van Nuenen, Tilburg University, Tilburg, The Netherlands

Rene van Stipriaan, Dutch Foundation for Literature, Amsterdam, The Netherlands

Puck Wildschut, Radboud University, Radboud, The Netherlands

Martin Wynne, Oxford University, Oxford, UK

Stefan Pernes, University of Würzburg, Germany

Steffen Pielström, University of Würzburg, Germany

Keli Du, University of Würzburg, Germany

Sina Bock, University of Würzburg, Germany