Research Networking Programmes

Science Meeting – Scientific Report

The scientific report (WORD or PDF file - maximum of seven A4 pages) should be submitted online within two months of the event. It will be published on the ESF website.

***Proposal Title**:* Exploring Historical Sources with Language Technology

Application Reference N°: 5618

1) Summary (up to one page)

The proliferation of digital resources in the Humanities is leading to the elaboration of new methods, concepts, and theories by means of which researchers can query and interpret large-scale textual collections. The goal of the workshop was to demonstrate how the application of language technology has produced a new understanding of texts in different fields of Humanities.The workshop brought together researchers who already apply language technology, and those who would like to learn about the current state of art in this new and evolving area. The organizers invited researchers (especially early career scholars) who plan to apply language technology but do not already have the necessary skills and technical background. The second main goal of the workshop was to enhance exchange of experiences, disseminate know-how, and to explore potential future collaborations.

2) Description of the scientific content of and discussions at the event (up to four pages)

Tony McEnery Lancaster University - *The Corpus as Historian - Using Corpora to Investigate the Past*

In this talk I will talk about work I have undertaken with historian Dr. Helen Baker looking at an issue in social history - prostitution in seventeenth century England. Social history in particular represents an interesting topic where the corpus might contribute - while the documentary sources and analyses associated with major historical events and figures are typically many and well analysed, the documents associated with the everyday, the unexceptional, are more sparse. In the case of marginalised or criminalised groups the documentary evidence outside of court proceedings is widely scattered and typically indirect. Prostitutes (we deal only with female prostitutes in this talk) are a good example of such a marginalised group - indeed in such a case the marginalization is enhanced by class and gender as well as criminality. I begin by considering what social historians have claimed about prostitution in this period. I then move to look at what the corpus shows us, using the latest version of the EEBO corpus available at Lancaster University - 1.5 billion words of lemmatised, POS tagged and spelling regularised written texts from the 15th, 16th, 17th and 18th centuries. Using corpus techniques to explore the texts, I show what a corpus may show the historian about prostitution in the period - and what historians can offer to corpus linguists who are approaching texts from this period.

Kat Gupta, University of Nottingham - *Constructing the "militant suffragist" in The Times*
This paper makes innovative use of critical theory to explore the media representation of the suffrage movement. The British women's suffrage movement was a complex, diverse campaign that emerged in the mid-nineteenth century. Focusing on The Times, I examine how suffrage campaigners' differing ideologies were conflated in the newspaper, particularly in connection with their support of or opposition to militant direct action. The effect of this reporting was to first conflate suffrage campaigners into a homogeneous mass and secondly to position all suffrage campaigners as proponents or supporters of militant direct action. Through this combination of established historical approaches, discourse analysis and corpus linguistic methodologies, this investigation refines our understanding of the suffrage movement in its socio-historical context and offers an insight into how language used by those in power can create and solidify identities imposed on fluid, polyvocal groups, particularly groups that lack access to their own media representation.

Carsten Schnober Technische Universität Darmstadt - *Welt der Kinder : Knowledge and Interpretation of the World as Portrayed in Textbooks and Children Books between 1850 and 1918*
The digital humanities project Welt der Kinder (”Children and their world“) started in May 2014 and is designed to serve as a template for similar projects in the future. By fostering close cooperation between historians, information scientists and computer scientists it aims to gain new insights into the period from 1850 to 1918; a time in which the accelerated production of knowledge was dominated by both globalization and nationalisation simultaneously. This material reflected contemporary world interpretation patterns and elements of cultural memory yet, equally, helped form the same. As its basic research resource, the project uses a continuously expanding corpus of several thousand books, which have been scanned and digitised using OCR technology capable of reading Gothic typefaces. The collection comprises more than 600,000 pages so far. We shall present the resources and methodologies applied in the project as well as its progress so far. We will discuss challenges, experiences, and problems encountered in the early stages of the project.

Stefano Menini, Fondazione Bruno Kessler - *Computational Analysis of Historical Texts*
In my presentation I will describe A.L.C.I.D.E. (Analysis of Language and Content In a Digital Environment), a system developed by the Digital Humanities group at Fondazione Bruno Kessler (Trento, Italy). A.L.C.I.D.E. is a web-based environment that uses human language technologies to help historians to extract, explore and analyze information within historical documents in real time. One of the goals of A.L.C.I.D.E. is to exploit these technologies by adapting them to the historical domain, and to make the results of linguistic and semantic analysis more interesting and usable by Humanistic researchers

through effective visualization techniques. As case study, we focused on the electoral debate between Nixon and Kennedy in the U.S. elections of 1960 (about 850 documents and 1.5 millions words). Our web platform allows researchers to browse and analyze the content of a document (or collection of documents), easily extract the needed information, and visualize it in a convenient interface, with the support of charts, dynamic graphs, maps, clouds of key-concepts and timelines.

Alex O'Connor, Trinity College Dublin - *Cendari: Leveraging Natural Language Processing for Research in Historical Archives*
An account of some of the challenges and successes associated with the Cendari Project's approach to leveraging Natural Language Processing and other Software Infrastructure to research in Historical Archives. Cendari takes a particular focus on collaboration between Information Experts, Technological Experts and Domain Experts. These different constituencies of users have significantly different perception of the available features, the potential value and perhaps even the merits of adopting NLP in Humanities Scholarship. Cendari is developing a set of flexible infrastructural services designed to support historical inquiry. This includes, from a natural language processing perspective, tasks such as information extraction for entity recognition, entity-driven search and annotation and sharing of research results. The project has manifested as different tools which are applied to user environments such as a Virtual Note-taking tool, and Archival Research Guides as well as support services for identity, access and provenance. The talk will include some details of specific results an findings, as well as overall results of how the process of joining different research communities has evolved.

Maarten van den Bos & Mariona Coll Ardanuy, Utrecht University & University of Trier - *Building a new political sphere? Early European Integration in Dutch digitized newspapers*
In the field of European Union studies, computational techniques to map out public discourse still are in their most early stage. A recent book on the role of national self-images uses a wide selection of more than a thousand editorials, but selection and analysis have been done merely by hand. Other studies that convincingly prove the investigatory value of newspapers also use traditional techniques to select and analyse the source material. By using the large repository of the Dutch Royal Library to extract social networks and their main topics of conversation, we will develop a method to further the history of European integration in a digital fashion.

Florentina Armaselu, CVCE - *Text Encoding and Enrichment for Linguistic Analysis: Archives on the policy of Armaments within Western European Union*
Using text encoding processes, the aim of the project is to explore text based methods and analysis to enrich digital data for exploration of the linguistic differences in the Armaments' policies of the Western European Union (WEU). The underlying research is driven by the need to explore the nature of the British and French positions on major security and defence questions in the WEU. The selected source material focuses on particular types of institutional documents associated with armament production and standardization. The objective is to take selected research material from the Archives Nationales de Luxembourg, WEU collection and use a process of data capture, enrichment and analysis to form new insight on the WEU corpus.

Tomaz Erjavec, Jožef Stefan Institute, Slovenia - *Modernising historical words*
The talk will present our results on modernising historical (Slovene) words, which enable better full text search in digital libraries and making old texts better understandable to today's speakers. Modernisation of word tokens also allows for applying annotation tools to be used on historical texts. We present the resources used in this research consisting

of a hand-annotated corpus, a lexicon of historical word-forms, and a large collection (digital library) of historical books. We then concentrate on the results obtained in our research on modernisation.

Ralf Morton Coventry University - *Using TEI mark-up and pragmatic classification in the construction and analysis of the British Telecom Correspondence Corpus*
The British Telecom Correspondence Corpus (BTCC) is a historical letter corpus which was constructed at Coventry University over a two year period from March 2012. The corpus contains a wide variety of business letters from the public archives of British Telecom, the world's oldest communications company. The BTCC offers a new way to explore this material and gain insights into the development of business correspondence over this period. Working with the TEI's Correspondence Special Interest Group I have devised a schema that makes it possible to filter the letters by a variety of contextual and linguistic categories depending on the individual researcher's aims. In this workshop I will present some general findings from the corpus as well as preliminary results from the functional analysis.

Susanne Haaf, Berlin-Brandenburg Academy of Sciences (BBAW) - *Text Type Classification for the Historical DTA Corpus*
This presentation focuses on the matter of text type classification for the Deutsches Textarchiv corpora. It will give insights about efforts and results of classifying DTA texts according to text types and the existing facilities for text type based research on the DTA corpus. Furthermore, current work on a new text type classification for the DTA will be presented. The goal of the project Deutsches Textarchiv is to create the basis for a reference corpus for the development of the New High German language (~1600–1900).

Renata Bronikowska, Polish Academy of Sciences - *Possibilities of searching the corpus of baroque texts – goals and objectives*
The electronic corpus of the 17th and the 18th century Polish texts (up to 1772) is being created in the Institute of Polish Language. It is meant to constitute a part of the National Corpus of Polish. The rules of transliteration and annotation of baroque texts ensure the achievement of two basic objectives: the faithful mapping of the notation and structure of ancient texts as well as providing the user with convenient methods of finding information in the corpus. By linking each word with an appropriate page identifier the search engine will provide the users with the ability to accurately indicate the location of the searched expression in the text, which will facilitate the use of quotations from the corpus in scientific works and dictionaries.

Victor de Boer, Netherlands Institute for Sound and Vision - *DIVE: Dynamically Linking Collections on the Basis of Events*
In this digital cultural heritage project, we provide innovative access to heritage objects from heterogeneous online collections. We use historical events and event narratives as a context both for searching and browsing as well as for the presentation of individual and group of objects. Semantics from existing collection vocabularies and linked data vocabularies are used to link objects and the events, people, locations and concepts that are depicted or associated with those objects. An innovative interface allows for browsing this network of data in an intuitive fashion. The main focus in DIVE is to provide support to (1) digital humanities scholars and (2) general audience in their online explorations.
The results from different tools and crowdsourcing are combined to come to high-quality extracted data. Target groups for evaluating the interface are Digital Humanities scholars, professional users and members of the general public.

Amelia Joulain, Lancaster University - *The spatial patterns in historical texts: combining corpus linguistics and geographical information systems to explore places in Victorian newspapers*
This paper reports on recent work that demonstrates that combining corpus linguistics and geographical information systems can help further our understanding of nineteenth century history. It presents an overview of work currently undertaken by the ERC-funded Spatial Humanities: texts, GIS, places project, whose aim is to develop and apply methods for dealing with textual data within a GIS environment. One example of our work, the Lake District project, analyses the historical cultural landscape of the Lake District by exploring a corpus of 80 texts (guidebooks, tourist notes and letters totalling >1,500,000 words). This research has demonstrated that accounts can be surprisingly different.

Lígia Gaspar Duarte, Évora University, Portugal - *Kinship: complex relation extraction in Portuguese eighteen century narrative sources*
Besides providing a general access to manuscripts, through digital technology it is possible to explore new tools in retrieval information, crucial fact to scholarly editions. Artificial intelligence work in natural language reveals interesting paths, although demands big efforts in domain knowledge representation.Trying to contribute to this effort, the present paper addresses the structure and automatic population form of the historical kinship ontology, for relation extraction in Portuguese early modern sources, as the early eighteen century printed and scribal news “gazettes” (1729-1742).

Maciej Eder, Institute of Polish Language of the Polish Academy of Sciences - *Stylometry and historical corpora: 8,281 Latin texts of the Church Fathers*
In the era of large-scale stylometry some basic but difficult methodological questions have not been answered yet. Certainly, the most important one is whether a given method, reasonably effective for a collection of, say, 100 texts, can be scaled up to assess thousands of texts without any significant side-effects. When one deals with historical corpora, however, this question becomes much more complex, since several additional factors have to be taken into consideration. Spelling variation, insufficiently trained NLP models, corpora a priori unbalanced – these are the obvious issues. The complete collection of Patrologia Latina, recently made available in the form of carefully prepared corpus with morphosyntactic annotation, gives us a great opportunity to test some of the above assumptions and possible drawbacks of the state-of-the-art stylometric methods. At the same time, however, the Patrologia Latina is a pre-internet example of a big-yet-dirty text collection. A number of massive stylometric experiments conducted on this collection partly confirm the aforementioned theoretical assumptions, but at the same time several new issues are revealed. This and similar results deserve a detailed linguistic (and literary) interpretation. This study is aimed at explaining some of the unexpected results.

Martijn Naaijer & Dirk Roorda, VU University & DANS - *Bible Research: Humanistic Information Retrieval*
The Hebrew Bible is a compact series of books. With its 426555 words it fits easily in your pocket, and with is 6 MB it fits many times in your smart phone. At the same time it is a body of texts shaped by people of different religious communities in varying geo-political circumstances over at least ten centuries of time. It is a complex cultural artefact, the object of intense research in religion, philology, history, and linguistics. In order to be able to use the methods of data analysis as they have been developed in recent years, it is important that the biblical text database exists in the open, in an accessible format, with well documented features, ready to be taken up by people coming from computational disciplines.In our presentation we will show some of the

results that are being gathered in and how this resource is a breeding ground where the above research questions can be tackled by a mix of computational philologists and philological computer scientists and many in between.

3) Assessment of the results and impact of the event on the future directions of the field (up to two pages)

In the two workshop days twenty-four short and long papers were presented that together showed a wide panorama of historical corpora, research questions and the digital tools that are used to enrich, query and analyse them. There is a vast number of digital texts and document collections from archives and libraries available for researchers from many different countries and periods. Especially the availability of (collections of) digitized newspapers has caught the attention of researchers who use them to retrieve opinions about all sorts of important events and developments. For earlier periods, newspapers are not available as a source, so researchers turn to a variety of different sources. Keynote speaker and linguist Tony McEnery in collaboration with historian Helen Baker studied views on prostitution in seventeenth-century England, using the large collection of digitized books of the Early English Books Online programme. He used a variety of computational methods to manipulate the book texts but pointed out that his *distant reading* was not meant as a form of 'culturomics' but mainly different approximations of getting an overview of what is in the texts and for using intuitions. Actual reading of (a selection of) the texts will always be necessary to get a proper understanding of the subject you are dealing with. In his words: 'close reading is the key'. The necessity of interaction between forms of exploration and quantitative analysis and refinement of questions and qualitative research was a result of especially the projects in which historians and linguists intensively collaborated. An interesting and slick example was presented by Victor de Boer who showed an interface from the DIVE project that enables users to explore the projects' texts, events and audio and visual materials and zoom in on them in different ways. Of course, there is no way of using linguistic or computational tools to get ready made results from a vast corpus of texts. Like all research, it is hard and intensive work.

The presenters did not propose one method as a favorite tool for text analysis. In fact, several presenters proposed to use method triangulation, that is the comparison of result of different methods of analysis and elaboration as the preferred way of coming to results. Most of the linguistic and computational tools in themselves were not very new; improvements in this field are gradual and incremental. It is not efficient to build a set of tools that exactly caters the needs of individual user (in this case a historian) or group of users. Historians also have to learn to apply the tools themselves to their own material and query and analyse the results themselves.

In the final discussion, it was noted that the presented work had many useful aspects for historians, but the emphasis on corpus linguistics did not quite reflect the historians way of working. They usually combines a variety of sources of which only a part is digitally available. And if they are digitally available, there is no digital text, because the originals are in manuscript and cannot all be transcribed of if there is digital texts, the computer recognized texts are hampered by poor accuracy. While linguistic methods can be a useful addition to the historians toolkit, their results will be combined with other methods, but they are worth the effort as the do enable historians to do distant reading and research of large amounts of texts that would not be conceivable using traditional methods.

4) Annexes 4a) and 4b): Programme of the meeting and full list of speakers and participants

Annex 4a: Programme of the meeting

Monday 8 December 2014

09:00	*Arrival, registration, welcome and introductions*
09:30	Tony McEnery & Helen Baker, *The Corpus as Historian - Using Corpora to Investigate the Past* (abstract, slides)
10:30	*Coffee break*
11:00	Kat Gupta, *Constructing the "militant suffragist" in The Times (abstract, slides)*
11:25	Carsten Schnober, *Welt der Kinder : Knowledge and Interpretation of the World as Portrayed in Textbooks and Children Books between 1850 and 1918 (abstract, slides)*
11:50	Stefano Menini, *Computational Analysis of Historical Texts (abstract, slides)*
12:15	Discussion
12:30	*Lunch*
13:30	Alex O'Connor, *Cendari: Leveraging Natural Language Processing for Research in Historical Archives, (abstract, slides)*
13:55	Maarten van den Bos & Mariona Coll Ardanuy, *Building a new political sphere? Early European Integration in Dutch digitized newspapers (abstract, slides)*
14:20	Florentina Armaselu, *Text Encoding and Enrichment for Linguistic Analysis: Archives on the policy of Armaments within Western European Union (abstract, slides)*
14:45	Discussion
15:00	*Tea break*
15:30	Tomaz Erjavec, *Modernising historical words (abstract, slides)*
16:00	Show and Tell Lightning Presentations *Presentations*
17:00	*End of day*
19:00	Dinner

Tuesday 9 December 2014

09:00	*Arrival*
09:15	Ralf Morton, *Using TEI mark-up and pragmatic classification in the construction and analysis of the British Telecom Correspondence Corpus (abstract, slides)*
09:40	Susanne Haaf, *Text Type Classification for the Historical DTA Corpus (abstract,slides)*
10:05	Renata Bronikowska, *Possibilities of searching the corpus of baroque texts – goals and objectives (abstract, slides)*
10:30	Coffee break
11:00	Victor de Boer, *DIVE: Dynamically Linking Collections on the Basis of Events (abstract, slides)*
11:25	Amelia Joulain, *The spatial patterns in historical texts: combining corpus linguistics and geographical information systems to explore places in Victorian newspapers (abstract, slides)*
11:50	Lígia Gaspar Duarte, *Kinship: complex relation extraction in Portuguese eighteen century narrative sources (abstract, slides)*
12:15	Discussion
12:30	*Lunch*
13:30	Maciej Eder, *Stylometry and historical corpora: 8,281 Latin texts of the Church Fathers (abstract, slides)*
13:55	Martijn Naaijer and Dirk Roorda, *Bible Research: Humanistic Information Retrieval (abstract, slides)*
14:20	*Tea break*
14:40	Discussion: *Language Technology and History,* respondents Jan Odijk and Rik Hoekstra
16:00	*End of workshop*

Annex 4b: Full list of speakers and participants

Marjolein	't Hart	Huygens ING
Lora	Aroyo	VU University Amsterdam
Miriam	Bouzouita	Ghent University
Renata	Bronikowska	Polish Academy of Sciences
Tessa	Carin Hauswedell	University College London,
Lígia	Clara Gaspar Duarte	Évora University, Portugal
Mariona	Coll Ardanuy	University of Trier
Max	De Wilde	Université libre de Bruxelles
Thierry	Declerck	University of Saarland
Katrien	Depuydt	INL
Marc	Dierikx	Huygens ING
Sebastian	Drude	CLARIN ERIC
Maciej	Eder	Polish Academy of Sciences, Institute of Polish Language
Bruno	Emanuel da Graça Martins	Instituto Superior Técnico, University of Lisbon
Tomaž	Erjavec	Jožef Stefan Institute
Kat	Gupta	University of Nottingham
Susanne	Haaf	Berlin-Brandenburg Academy of Sciences and Humanities (BBAW)
Harald	Hammarström	MPI Nijmegen
Lex	Heerma van Voss	Huygens ING
Ben	Heuwing	Hildesheim University
Rik	Hoekstra	Huygens ING
Amelia	Joulain	Lancaster University
Dimitrios	Kokkinakis	University of Gothenburg
Rafał	L. Górski	Institute of Polish Language, Polish Academy of Sciences
Tony	McEnery	Lancaster Unversity
Stefano	Menini	Fondazione Bruno Kessler, University of Trento
Ralph	Morton	Coventry University
Jan	Odijk	Utrecht University
Mike	Olson	Utrecht University
Katie	Patterson	University of Liverpool
Wim	Peters	University of Sheffield
Ana	Pyltowany	University of Amsterdam
Claudia	Resch	Institute of Corpus Linguistics and Text Technology, Austrian Academy of Sciences

Rosa	Ricci	University of Leipzig
Dirk	Roorda	DANS
Federico	Sangati	FBK, Trento
Mari	Sarv	Estonian Literary Museum
Antoinette	Schapper	KITLV, Leiden
Carsten	Schnober	German Institute for International Educational Research
Christof	Schöch	Würzburg University
Mari	Smits	Huygens ING
Gabor	Toth	Passau University
Karina	van Dalen	Huygen ING
Maarten	van den Bos	Utrecht University
Adam	Wyner	University of Aberdeen
Martin	Wynne	University of Oxford
Raphael	Zahnd	University of Zurich