REPORT

Workshop 21 November 2012

‘Using Large-Scale Text Collections for Research: Status and Needs’

First workshop of the NeDiMAH working group Using Large-Scale Text Collections for research.

Location: Huygens ING, The Hague, The Netherlands
Organizer: Prof. dr Karina van Dalen-Oskam

1. Summary

1.1 Summary of the aim of the working group Using large-scale text collections for research

IT tools and methods, such as information retrieval and extraction methods (including, for example, text and data mining), can reveal new knowledge from large amounts of textual data, extracting hidden patterns by analyzing the results and summarizing them in a useful format. This working group will examine practices in this area, building on the work of corpus linguistics and related disciplines to develop a greater understanding of how large-scale text collections can be used for research.

1.2 Summary of the 1st workshop: Using large-scale text collections for research: status and needs

The meeting was used to assess the availability of text corpora for researchers from different disciplines in the participating countries and languages. How large are the available corpora? For what purposes were they created? What kinds of mark-up do they contain? And which tools are available to help mining the corpora? What is missing in both texts and tools to make the corpus also useful for other research disciplines than the one it was originally created for? After a general introduction by the working group leader, three papers were presented followed by discussion. After this, the other participants sketched the situation in their country and language(s) and the needs of their own specific research discipline. Most of the afternoon was spent on discussions about the topics addressed during the first parts: what are the shared positive points in the different countries/languages/disciplines? Is there an overlap in the different needs that were expressed? What can we learn from each other? Where can we push the developments further through a shared approach? At the end of the day, the participants phrased the main points that were raised and listed possible next steps, such as the topics of the next workshops and/or seminars to be organized by this working group.

2. Description of the scientific content of and discussion at the event

Adam Kilgarriff demonstrated his sketch engine software and explained the business model behind his approach. The basis of most functionality is creating smart concordances. The corpora in the Sketch Engine are either open access or limited in access. This was the start of the recurring discussion on the topic of "the right to be forgotten" on this day, in this case the intricacies of intellectual property rights of the included corpora. Kilgarriff stated that his business model is a longer term guarantee of sustainability than academia.

Duncan Berryman presented his research into medieval buildings. He collects his information in a rather elaborate Excel file, with such columns as prices for building as known from medieval charters and registry books. He demonstrated how widely the kinds of information are in his research. It became clear that there is a big need for an easy to use (online) database for data gathering such as this, but those attending the workshop could not provide Berryman with adequate suggestions in this area. His presentation also leads to the identification of the pressing need for good OCR software for medieval material. If that will ever be possible, researchers such as Berryman could include much more material during the short duration of their research project. Katrien Depuydt referred to the outcomes of the IMPACT project, in which OCR possibilities have been enhanced. She will send information about the center of expertise that resulted from the IMPACT project later.

Andreea Popa showed how many changes totalitarian regimes have brought about in urban landscapes in the last century. She showed a web portal in which many publications on this topic are available, but all in non-machine-readable pdf format. In practice, this means that her project follows the same procedure as Berryman already demonstrated: visual inspection of digital material leads to the inputting of data in e.g. a database or a scholarly paper. Since most of the material in the portal is from the 20 century, the IPR is a big problem as well. During the discussion it became clear that there is a big difference in digitization projects on a national level. There are hardly any national initiatives in Rumania, which means that scholars are still severely limited in what they can use, and have to decide for themselves if they will digitize materials themselves or not. In Croatia it seems that digitization is only done as a kind of window-dressing, and not yet for serious scholarly use. There is a clear need of knowledge as has been generated in the IMPACT project already mentioned. The same remarks were made when asked whether GIS systems are in wide use in Rumania.

Pim Huynen talked about examples of mining a corpus of texts, using word clouds to trigger ideas, as an exploration tool, or timelines. He wonders, however, whether these kind of tools are really useful for a scholar, since it is extremely unclear what the results exactly can tell us. Many colleagues have this skeptical approach. Others agreed with Huynen that many scholars have a fear of the new methods and means. It does lead to attempts to develop new tools to reduce the amount of noise in hits of tools. Huynen also showed an example of sentiment mining in relation to certain concepts in different languages, and emphasized that comparing results across languages is extremely problematic.

Miguel Costa started his short talk by stating that we are now living in the digital dark ages. He then demonstrated his work on the project that archives Portuguese web pages. The project makes uses of tools such as language recognition and detection of new pages through links. He described that Portugal does not have many limiting IPR rules yet. When somebody complains about the archiving of a certain page, it is removed.

Thomas Eckart gives a frightening insight in his work at the Leipzig Corpora Collection. It collects modern material mainly based on the web. Many languages from central Europe are included, which sometimes is a problem where no one in the projects can recognize what language a text is written in. The project programmed a set of web services, which help users to answer relatively (technically) simple questions easily. Eckart described the IPR problems related to his project, which includes the republication of i.e. newspapers. He referred again to “the right to be forgotten”. When people want to not be mentioned in the corpus, their names are put on a list of names and the sentences in which they occur to be removed from the corpus. He pointed out that a closed infrastructure such as aimed at by CLARIN might be one way of solving this problem.

Ulrike Henny briefly sketched her work on a TEI edition. She addressed the need for more tools to help mine TEI documents, and asked what corpus builders need of a digital text edition to make the edition a worthwhile contribution to a corpus. What formats to use? What export options to offer? How can an edition be automatically harvested by a corpus? In the discussion her questions were expanded from technical questions only to questions into the chosen transcription type as well, since linguistic corpora have other needs than a broad reading audience of an edition, leading to a preference for diplomatic transcriptions instead of respelled reading texts.

Gordan Ravancic described the lack of funding for digital humanities projects and digitization in Croatia. Institutes doing anything in these areas do not collaborate at all, which means that there is no cohesion whatsoever on a national level. TEI is very important but mainly in linguistics, not in e.g. historical research. Ravancic demonstrated several small projects with interesting approaches, i.e. the tool CatViz, by Artur Šilić of the faculty of electronics and robotics of the university of Zagreb, which helps to analyze text based on positive and negative adjectives.

Katrien Depuydt gave a very clear description of the workflow of lexicographers, and how one of the important elements in the digital workflow is to differentiate between the source text and the editor text. Milestones are needed fir this. She also went into the innovations of the IMPACT project on OCR more deeply. A problem that was pointed out by her and which seems to be relatively recent is that different versions of complete corpora are turning up, which seems to be rather confusing and could lead to severe problems.

3. Assessment of the results and impact of the event on the future direction of the field

During the workshop the following main topics were dealt with and agreed on as being the main topics by all:

1. There is an agreement on that all have a need for large corpora of texts, but that the situation in the represented European countries differs extremely as to already available digital texts and digital corpora. A sub question was: what is large? Our preliminary definition is: A corpus that you can only use well if you make use of software / have to use a quantitative approach.

2. All participants have a need for smart tools for several kinds of use and analysis of large text collections.

3. Everybody has some form of problems when it comes to legal aspects - what is allowed in a digital corpus and what has to be removed when people ask for that? (The problem of "The right to be forgotten").

4. The status of available machine-readable text through OCR (Optical Character Recognition) differs enourmously across Europe. Information about the IMPACT project, which resulted in a Center of Competence for OCR, will be distributed through the more detailed report to follow.

5. There is a need for help in learning to deal with mass data and how to evaluate and interpret them (heuristics).

5. Scholars need help in convincing other scholars about the usefulness of the use of large text collections by using statistical analysis etc. (How can we bring about a change in methodological thinking in our research disciplines?)

6. Convincing scholars is only possible when we are able to explain the ‘black boxes’ of all the new tools, for which we needs new kinds of expertise (by collaboration with the right specialists). In this whole context, statistics becomes very important again. There is a generally positive feeling on the new research possibilities - although some experience rather aggressive feedback when they actually do this kind of research.

One of the ideas is to organize the seminar as some kind of tools market. Most of the participants were willing to contribute or co-author a chapter in a book or an article. All would like to be kept in the loop of the working group's progress.

The working group leader is currently preparing a detailed proposal for the next steps.

4. Final programme of the meeting

9.00 – 9.15: Welcome with coffee / tea

9.15 – 9.45:
Karina van Dalen-Oskam (Huygens ING / University of Amsterdam), **Introduction: aims and tasks**

9.45 – 10.15:
Adam Kilgarriff (Lexical Computing Ltd., United Kingdom), **The Sketch Engine as Infrastructure for Large Scale Text Collections for Humanities Research**

10.15 – 10.45:
Duncan Berryman (Department of Archaeology & Palaeoecology, Queen's University Belfast, United Kingdom), **Archaeology from Documents: Using large collections of accounts to investigate medieval buildings**

10.45 – 11.15: Coffee break

11.15 – 11.45:
Andreea Popa (Department of Urban and Landscape Design (DULD), Faculty of Urban Planning (FUP), University of Arihitecture and Urbanism Ion Mincu Bucharest (UAUIM), Romania), **Use of large scale text collections in order to reveal Urban Landscape History**

11.45 – 12.15:
5-minutes' presentations and discussion of other attendees about their involvement in usage of large-scale text collections, describing their experiences, problems, and wishes

12.30 – 13.30: Lunch (together with the NeDiMAH working group Digital Scholarly Editions)

13.30 – 15.00:
5-minutes' presentations and discussion (continued) of other attendees about their involvement in usage of large-scale text collections, describing their experiences, problems, and wishes

15.00 – 15.30: Tea break

15.30 – 17.30:
General discussion, planning of next steps in the NeDiMAH working group

17.30: Drinks (together with the NeDiMAH working group Digital Scholarly Editions)
19.00: Dinner at a local restaurant

List of attendants

Duncan Berryman	Queen's University Belfast, Belfast, Northern Ireland	dberryman01@qub.ac.uk
Miguel Costa	Foundation for National Scientific Computing, Portugal	miguel.costa@fccn.pt
Katrien Depuydt	Institute for Dutch Lexicology, Leiden, Netherlands	Katrien.Depuydt@inl.nl
Thomas Eckart	NLP group University of Leipzig, Germany	teckart@informatik.uni-leipzig.de

Ulrike Henny	Cologne Center for eHumanities (CCeH)	ulrike.henny@uni-koeln.de
Pim Huijnen	Utrecht Institute for Pharmaceutical Science, Utrecht, Netherlands	p.huijnen@uva.nl
Adam Kilgarriff	Lexical Computing Ltd., United Kingdom	adam@lexmasterclass.com
Andreea Popa	University of Arhitecture and Urbanism Ion Mincu Bucharest, Romania	apopa.uauim@gmail.com
Gordan Ravancic	Croatian Institute of History, Zagreb, Croatia	gordan@isp.hr
Karina van Dalen-Oskam	Huygens Institute for the History of the Netherlands, The Hague, Netherlands	karina.van.dalen@huygens.knaw.nl