Florilegia: Big Textual Data Workshop,
July 10-11, 2017

Final Report
(October 2017)

Universität Leipzig
Augustusplatz 10-11 – 04109 Leipzig

Within the framework of the BMBF funded *Global Philology Planning Project*, Dr. Thomas Koentges, University of Leipzig, has organized a workshop on *Big Textual Data*.

The workshop was held in the Institute of Computer Science at the University of Leipzig on July 10-11, 2017. The program of the workshop is available through this link: http://www.dh.uni-leipzig.de/wo/events/global-philology-big-textual-data/ . Attendees and speakers tweeted the event with the hashtag **#GPFlorilegia2017**. Tweets are available thorugh this links:
https://twitter.com/hashtag/GPFlorilegia2017?src=hash

The *Florilegia: Big Textual Data Workshop* focused upon automated methods of analysis that extend individual cognitive power by applying complex computational philology methods, tools, and services. The workshop focused on three aspects of complex computational philology: analysis, synthesis, and abstraction.

Regarding the analysis, questions of scale and tokenisation: “How big is big?” & “What are my tokens (e.g. words, character-ngrams, syllables, quantities)?” were discussed, but also methods of data preparation and transferability from one language (or method) to the other: “how do we analyse the tokenized corpus or multiple different tokenisations of the same corpus” and “which method works for which tokenisation”? When discussing synthesis, we concentrated of questions like “how to combine the results of different analyses meaningfully?” and “how to avoid circular arguments (or biasing your data) when performing multi-method complex computational philology?” And finally, regarding the abstraction we discussed methods of meaningful reduction. Multi-method big textual data research creates a lot of output, but how can we express our n-dimensional data meaningfully? How can we communicate our research results.

15 speakers were invited to participate in the workshop. The speakers belonged to institutions of 4 different countries (Croatia, Germany, the Netherlands and the United States of America) researching 7 languages (Arabic, Chinese, Coptic, English, Greek, Latin, Sanskrit) of 3 different language families..
The two days of the workshop had three sessions each day, including opening and closing discussions.

July 10, 2017
After welcoming speakers and attendees and after the introduction to the workshop by **Thomas Koentges** and **Gregory R. Crane**, the first part of the morning was devoted to the

presentation of projects and research problems pertaining to traditional questions of text reuse.

David Smith *Exploiting Relational Structure in Large Text Corpora* gave the opening talk demonstrating big data methods applicable for all languages. This was followed by two **how-to** presentations: **Benjamin Kiessling** presenting on Leipzig's OCR workflow and algorithms for all languages & Alicia Gonzalez explaining how to push annotations of different languages to *Annis*.

The next session dealt with **Deep Learning and Topic Modelling**, starting with **Oliver Hellwig** presenting on a deep learning approach to tokenise Sanskrit, followed up by **Paul Dilley** and **Thomas Koentges** who had produced several topic models for the Iowa Canon of Latin texts (30 million tokens).

In the end of day discussion, it the transferability of the presented methods were discussed, mainly related to tokenisation challenges in different languages regarding morphology and token frequency.

July 11, 2017
The morning started with a session on **Corpus Infrastructure and Resources**. **Thomas Koentges** demonstrated the use-case for a new data format called .cex, **Frederik Baumgardt** showed how to robustly anchor annotations to changing text, and **Patrick J. Burns** presented on the Classical Language Toolkit, a python library collection to process historical languages.

After a coffee break the participants welcomes some talks about **Corpus Building**:
Cliff Wulfman showed the challenges of building and making accesible a corpus of modernist and avant-garde magazines, **Matt Munson** demonstrated use-caseds for our CTS API and the first one-thousand years of Greek corpus, **Neven Jovanović** introduced us to the Croatiae Auctores Latini (CroaLa), a Neo-Latin Corpus, and finally **Tyler Neill** spoke about a digital critical edition of the **Nyāyabhāṣya.**

In the last panel of the day, **Big Textual Data and Text Reuse**, **Donald Sturgeon** and **Paul Vierthaler** presented their approaches to massive Chinese language corpora,

The concluding discussion dealt again with questions and solutions regarding the transferability of the methods discussed and similarities of individual language features and general infrastructure was seen as an advantage of Global Philology. The participants continued to discuss topic of the workshop in the slack-chat that was established for the conference: https://florilegia.slack.com and wished that the event would be of recurring nature.