Processing Ancient Text Corpora

2020-20-17/21 Lorentz Center@Oort/Snellius

Scientific

Description and aims

This workshop aims to promote scholarly exchange and to build a community of scholars with an interest in digital humanities and ancient texts. Research into ancient texts undergoes strong development: the application of ever more methods from statistics and machine learning. Given the fact that the text disciplines are organized by language, rather than by method, we think the methodological exchange can be strengthened. The sharing of IT techniques is a natural playground for this, but only a starting point. Theoretically, we need to discuss where these methods bring us. Are big data methods also applicable to small data? What is particular about the fact that the texts of interest are historical? Practically we want to discuss how we can optimally employ IT methods. Can we assess the landscape of IT and make an informed selection of the regions that are most useful to us?

Tangible outcome

We found two aspects of text processing at opposite ends of a spectrum that deserve closer interest:
Material: ancient text corpora have a very material dimension: manuscripts, tablets of clay, inscriptions in bowls. Text should not only be annotated with linguistics and interpretation, but also with material traits.
Interface: many classical are stored in the Perseus Digital Library, from where they are accessible in an online reading environment, the Scaife Viewer. It does not accommodate intensive processing but it could support Text-Fabric import and export, so that results of computing can be brought back to the reading environment. One of the participants has already written code that can convert texts in the Perseus library to TF. Two others are overseeing the renovation of the Scaife. There was definite interest to develop TF-interoperability in Scaife.

Breakthrough

A problem that all participants gave a good deal of thought was: how exactly can philologists adopt computing as a tool in their research? Should they partner with web developers and data scientists? Should they become software engineers themselves? Will *research software engineers* be able to wade through the huge amounts of domain-knowledge of philologists before becoming actually useful?
Machine learning is great, but it cannot be applied to philological data without careful consideration, and it needs conscious effort to make sense of the results. Corpus-driven software for language learning is an art of its own, which can help scholars to make sense of corpora in languages outside their expertise.
Then it dawned upon us that processing ancient corpora is a skill on its own. We need to identify as a guild, in order to develop the tools of our trade and instill them in new students. Instead of soldiering on with ad-hoc IT solutions, we must find our way competently in the vast array of IT tools and frameworks.
We will reinforce the competence of a computing philologist. We need to anchor it in organizations, and if needed, we will create a new one, such as a Text-Processing Foundation.

"Aha" moments

When dealing with historical texts, there is so much more information than the bare text: the context. Moreover, philology is interested in the human mind as it leaves traces in the material world. When processing texts, it is a challenge to make use of that information. It is important to collect and represent it in knowledge organization systems, such as the Semantic Web, rather than tying it too closely to a corpus.

Organization

Format of the workshop

The workshop was conducted in an informal way. We did have lectures, but also sessions for hands-on work and discussions. However, during the workshop it turned out that many attendants took the opportunity to convene in small groups to work on the issues at hand. The hands-on sessions were too short to let everyone set up his/her computer properly. The planned discussion sessions gradually gave way to spontaneous work in smaller groups.

Comments

Next time we will reserving more time for hands-on sessions, effectively shifting to a boot camp setup. It remains important that participants can select and develop topics during the event.
Certainly, the current accommodation, with one big space and several smaller spaces, would also be well-suited to such a format.

Wido van Peursen (Amsterdam, Netherlands)
Nicolai Winther-Nielsen (Copenhagen, Denmark)
Cody Kingham (Cambridge, United Kingdom)
Dirk Roorda (The Hague, Netherlands)

Presentations archived at ZENODO 10.5281/zenodo.3719091