Processing Ancient Text Corpora

Twenty five scholars and profesionals working with ancient texts have spent a week at the Lorentz Center Leiden in order to discuss the state of the art of researching ancient text corpora. It was not only about research, but also about better ways to teach domain knowledge: ancient languages, history, and archaeology. It was not only about domain knowledge, but also about the expertise of employing digital tools for the jobs, and how to transfer it to every new student generation and to ourselves.

Colofon

17 - 21 February 2020

Venue: Lorentz Center@Snellius

Link to the conference

Link to the presentations

Organizers and participants

- Nicolai Winther-Nielsen, FIUC-DK & Vrije Universiteit
- Dirk Roorda, Data Archiving and Networked Services
- Wido van Peursen, ETCBC, Faculty of Religion and Theology
- Cody Kingham, Cambridge University

Last name	First name	Institute	City	Country
Bhulai	Sandjai	Vrije Universiteit	Amsterdam	Netherlands
Bleeker	Elli	R&D group, Humanities Cluster, KNAW	Amsterdam	Netherlands
Boogert	Ernst	Protestant Theological University	Amsterdam	Netherlands
Coeckelbergs	Mathias	Université libre de Bruxelles / KU Leuven	Brussels / Leuven	Belgium
Crane	Gregory	Tufts University	Medford	United States
de Ridder	Alba	Universiteit Leiden	Leiden	Netherlands
Erwich	Christiaan	ETCBC, Vrije Universiteit	Amsterdam	Netherlands
Fekadu	Zetseat	Wycliffe	Addis Ababa	Ethiopia
Folmer	Margaretha	ETCBC, Vrije Universiteit	Amsterdam	Netherlands
Glanz	Oliver	Andrews University	Berrien Springs	United States
Haentjens Dekker	Ronald	R&D group, Humanities Cluster, KNAW	Utrechht	Netherlands
Højgaard	Christian Canu	FIUC-DK & Vrije Universiteit	Copenhagen	Denmark
Johnson	Cale	University of Birmingham	Birmingham	United Kingdom
Kingham	Cody	Cambridge University	Cambridge	United Kingdom
Paulus	Erick	LIACS, Universiteit Leiden	Leiden	Netherlands
Robar	Elizabeth	Tyndale House	Cambridge	United Kingdom

Last name	First name	Institute	City	Country
Roorda	Dirk	Data Archiving and Networked Services, KNAW	Den Haag	Netherlands
Sikkel	Constantijn	ETCBC, Vrije Universiteit	Amsterdam	Netherlands
Talstra	Eep	ETCBC, Vrije Universiteit (em.)	Amsterdam	Netherlands
Tauber	James	Eldarion	Boston	United States
van Hecke	Pierre	KU Leuven	Leuven	Belgium
van Lit	Cornelis	Utrecht University	Utrecht	Netherlands
van Peursen	Wido	ETCBC, Vrije Universiteit	Amsterdam	Netherlands
Winther-Nielsen	Nicolai	FIUC-DK & Vrije Universiteit	Copenhagen	Denmark
Wyns	Roxanne	LIBIS, KU Leuven	Leuven	Belgium

Corpora

Several corpora were represented by people present that have worked with them. Here is an overview:

Category	Language	Specifics	People involved
Perseus Library	Greek	ancient	Gregory Crane, James Tauber
Perseus Library	Greek	patristic	Ernst Boogert
ETCBC Bible	Hebrew, Syriac	Leningradensis, Dead Sea Scrolls, Peshitta	Eep Talstra, Wido van Peursen, Constantijn Sikkel, Christiaan Erwich, Pierre van Hecke, Matthias Coeckelbergs, Oliver Glanz, Christian Højgaard, Cody Kingham, Nicolai Winther-Nielsen
Inscriptions	Hebrew	Magic bowls	Margaretha Folmer
Inscriptions	Old Sundanese	Palm leaf corpus	Erick Paulus
Medieval Islamic	Arabic	Ibn 'Arabi	Cornelis van Lit
Cuneiform Tablets	proto-linguistic, Akkadian	Uruk, Old Babylonian, Old Assyrian, etc.	Cale Johnson, Alba de Ridder
Native speaker Aramaic	Neo-aramaic	NENA, Cambridge	Cody Kingham

Digital tools

Category	Systems/Software	Projects	People involved
Infrastructure	Library system, ontologies	ReiRes	Roxanne Wyns
Linked Data	Semantic Mediawiki	Onomastics	Elizabeth Robar

Category	Systems/Software	Projects	People involved
Sandjai Bhulai	Machine Learning, Neural Networks, TensorFlow	Pilot: Biblical Hebrew POS tagging with Markov models and neural networks	
Data models for text	Text-As-Graph	Alexandria	Ronald Dekker, Elli Bleeker
Data models for text	Text-Fabric	annotation	Dirk Roorda, Ernst Boogert, Cody Kingham, Oliver Glanz, Christian Højgaard, Cale Johnson, Alba de Ridder
Web interface	Reading environment	Scaife Viewer	James Tauber, Gregory Crane
Image processing	OCR, OpenCV	opencv	Cornelis van Lit

Teaching

Category	Systems/Software	Projects	People involved
Teaching aids	Bible Online Learner	DADeL, ...	Nicolai Winther-Nielsen, Oliver Glanz, Ernst Boogert, Zetseat Fekadu
Image processing	OCR, OpenCV	Among Digitized Manuscripts	Cornelis van Lit
Tutorials	Text-Fabric	notebooks	Dirk Roorda
Tutorials	SHEBANQ, Text-Fabric	video	Oliver Glanz
Tutorials	Text-Fabric	notebooks	Christian Højgaard

The overall picture

The keynote lecture by Gregory Crane and James Tauber offers a concise description of the current situation of philology: its mission, challenges, opportunities, and choice of methods.

In a nutshell, philology is concerned with the textual record of man-kind for the sake of studying the circumstances of people and the events that happened to them. It borders on archaelogy, but it is specialized to the interpretation of symbols on artefacts, especially when those symbols are decipherable, and especially if the decipherment is linguistic text. Given this goal, philology is not wedded to a particular method of research, as long as the results remain firmly connected with the surviving materials.

Historically, actual philology has had periods in which it became bound up with theology, in the study of sacred texts. This was put to an end in the 19th century by Friedrich Wolf when he delineated the science of Antiquity (Altertumswissenschaft).

This amounts to a threefold drive to integrate the many philological subdisciplines:

1. all of them need to re-conceive their methods in view of what is digitally possible;
2. there is no methodological separation between secular and sacred texts;
3. all texts of Anitiquity are needed for the full picture, not just Greek and Latin.

Building out: grants, a foundation?

On the last day of the conference we exchanged thoughts about consolidating our work. The creators of some of the techniques and models and apps we have seen in this conference are predominantly past half-way their career, the adopters are in majority well before that point. We need to establish continuity between idea, pilot, tool, and classroom materials.

One of these ideas could be the following.

Text Processing Foundation (TPF)

An expertise center for the processing of cultural heritage texts. Covering

- Software - how researchers can interact with all the tools available and adapt them for the sake of processing texts from the past;
- Data - how the texts can best be represented for the various purposes of editing, processing, archiving, and showcasing;
- Infrastructure - how researchers can cross the borders of their own laptop in order to gain more processing power, data storage, and sharing options

The emphasis is more on the people aspect than on the machinery aspect. Because: the biggest challenge is the fragmentation of skills between the many humanities disciplines. We need to bundle them and make those skills adoptable by people with a deep focus on their subject material.

The scope of interest is everything from the material artefacts (bordering on archaeology) to advanced digital processing of patterns (bordering on artificial intelligence) and all the steps in between.

A role model is the Software Sustainability Institute in the UK.

The position could be a nationally funded initiative with international contributors. In the Netherlands it should have close links with the eScience Centre, the Humanities Cluster, DANS. There could be commercial partners as well, e.g. Brill and Triply. And there should be an interface on ongoing infrastructural humanities projects such as CLARIAH.

The results of the TPF over time

- Summer boot-camps (Lorentz-like conferences but with hands-on work in the foreground);
- Development of reference data, and turning it into Linked Data: Elizabeth's work (bottom up data definitions) plus Roxanne's work (integration of semantic data in the Linked Data);
- Research Software Engineering in the middle of Linked Data, Corpora, Interfaces, Machine Learning; empower corpus analysis tools with built-in support for sharing and reusing data and making use of linked open data;
- Expertise building in advanced analysis techniques: adaptations to machine learning, selections of machine learning so that it works on small corpora;
- Expertise building in digitisation, leading to workflows from the ur-digitisations of the texts to intermediate, processing-ready representations of the texts;
- Historical Linguistics: nearly all text is linguistic by nature. We need NLP tools that can work on historical texts.

This is only possible if the TPF operates in an international setting, as one of the players. The TPF should build on TEI, work together with builders of interfaces such as the Scaife Viewer, and with teams that rethink what digital editions should look like.

Reflection

During the conference we became aware of several characteristics of the current state of philological research. It dawned upon us that quite a few of us work at the edges of our home institutions, doing things that are sometimes easier shared with like-minded spirits far away than with nearby colleagues.

Another observation is that the philologists among us tend to work with our corpora in different ways than they have been taught. They do not only read them closely, but also want to query them, mine them, and interlink them, using tools that they are just about able to manage. It is difficult to convince (some of) their peers that these tools are worth the hassle.

The ones who are engaged in delivering computing solutions to the philologists discover that the regular tools of the trade are often ill-adapted to the use cases in philology. Quite some ingenuity is needed to work around limitations, let alone to develop brand new solutions that have the needs of philologists in focus. One factor is that the data of philologists is often surrounded with layers of expert knowledge which can be very opaque to the untrained eye.

At the same time we see a steady rise in the readiness of philologists to adopt IT. This goes hand in hand with the increased user friendliness of programming. Python and its ecosystem of libraries for scientific computing and visualization has brought programming to researchers who are not primarily trained in computer science. Especially Jupyter notebooks help to present computer programs as narratives that can actually be read by a human.

The take home messages

At the end of the conference the participants voiced there take home messages:

Constantijn Sikkel: the TAG model is interesting, and there is an interesting challenge to treat an Akkadian corpus with the methods used for the Hebrew Bible and then see to what extent we can reuse knowledge of Hebrew for Akkadian.

Elizabeth: the long-term effects of the ontologies we develop. It's worth the extra time to research others' work, for the sake of scholarship at large, rather than thinking only of the needs of our particular project.

James Tauber: there are ways of integrating Scaife and TF, and there is also a definite interest to use such an integration.

Roxanne Wyns: it gave me a better picture of where data is created, by which tools, and how that can feed in in further research.

Cody Kingham: encouraged to stay deeply immersed in technology.

Dirk: we are developing an identity as researchers computing with ancient texts.

James (for Greg): thrilled by the work of a community that I have not been deeply involved in and seeing lots of convergence.

Oliver Glanz: many benefits from 1-1 conversations during this conference.

Zetseat Fedaku: I'm not only lame but also blind🤓 . Got a better picture of how digital humanities are relevant for my work (especially Cody's presentation).

Cale Johnson: the realization that data can move between different systems while keeping structure.

Wido van Peursen: previous Lorentz Workshop in 2012 (link forthcoming). But a thing like shebanq was not cooked at that meeting itself, but a few months after the workshop the connections were made to write a proposal to a CLARIN call.

Hopefully that holds for this conference too: we are not in control but we can plough the fields and do some sowing.

Cornelis van Lit: next steps needed for ancient text research driven by scans and photos (LWCvL):

Programmatic access (of catalog and photos) Photos of all parts of the artifact No watermarks Holding information in the filename Allowing results back into the catalog Standardized way of describing and sharing features