REPORT

On the 6th and 7th of July 2017 the *Linguistic Annotation and Philology Workshop* has taken place at Leipzig University, Department of Computer Science (Digital Humanities section). The aim of the workshop was to gather scholars working within the field of Philology, Linguistics, Digital Humanities, and Computer Science to survey the status of the art concerning linguistic resources for historical languages, so as to plan concerted lines of intervention. Twenty-five scholars took part in it, coming from ten different (EU and non-EU) countries.

The ability to computationally access and analyze historical languages is acknowledged to be a key component to study the ancient world: written testimonies are a/the major source of knowledge of the past, and so linguistic and non-linguistic studies heavily rely on them for information retrieval. To this end, digitalization and annotation of ancient documents, which scales up and can be relied on, is paramount to continue their *traditio* ('transmission') in the digital age. Scholars acknowledge the need to correctly address the challenges of translating and annotating ancient documents posed by the digital environment, which their successful preservation and query (and so knowledge) crucially depend on.

Application of the latest technologies to parse many non-mainstream historical languages is still, in many respects, in its infancy. The detailed program of the conference and the abstracts/brief summaries for each presentation follow at the end of this document. In general, the participants acknowledged the importance of the following points:

- One should rely on already existing frameworks, avoiding dispersion with building new, *ad-hoc*, resources. This is due not only because building on existing resources is usually more affordable, but also because such resources are usually more sustainable on the long run. Community efforts concentrate on them more easily and so their future is more reliable. Likewise, if a new resource has to be started, this should be generalized as far as possible.
- Communities should use standard technologies as far as possible. Some examples include TEI (text encoding initiative), UD (Universal Dependencies), or, more in general, technologies suchs XML o JSON for data exchange.
- One should rely on already existing platforms allowing scalable data creation and sharing, such as, for example, GitHub and CLARIN.

WORKSHOP PROGRAM

09:00-09:30
Giuseppe G. A. Celano & Gregory R. Crane, Universität Leipzig (Leipzig University) & Tufts University
Introduction

09:30-10:00
Timo Korkiakangas, Universitetet i Oslo (University of Oslo)
Quantifying spelling variation: Scribes' command of Early Medieval documentary Latin

Abstract
This talk gives an example of how a morphologically annotated treebank can be utilized beyond what the corpus was originally intended for. I will investigate whether the considerable oscillation in spelling in the Latin of early medieval private documents (charters) correlates with the oscillation in employing certain morphosyntactic categories. My hypothesis is that the more non-standard the spelling is, the more frequent are the novel, Romance-type, morphosyntactic constructions and the poorer the command of those Classical constructions that were in decline in Late Latin. Spelling variation is here operationalized by normalizing the non-standard word forms into standard-Latin forms and subsequently quantified by calculating the edit-distance between all the word forms of the corpus and their standard-Latin counterparts, whether these be originally standard or non-standard. The normalization of word forms is based on the two-million- word Open Office Latin lexicon which I lemmatized and tagged morphologically with Whitaker's WORDS tagger. The resulting word form library consists of one or more lemma/morphology entries for each word form. This library was then used to produce Classical-Latin forms out of the lemma/morphology pairs attached to each word in the Late Latin Charter Treebank (LLCT, 480,000 words), which consists of 1,040 charters from Tuscany of the 8 th and 9 th centuries.

10:00-10:30
Camilla Di Biase Dyson & Simon Schweitzer, Georg-August-Universität Göttingen (University of Göttingen) & Berlin-Brandenburgische Akademie der Wissenschaften (BBAW)
Stand-off Annotation to Ancient Egyptian text corpora based on BTS

Abstract
The Thesaurus Linguae Aegyptiae (TLA; http://aaew.bbaw.de/tla) is the publication platform of the project „Structure and Transformation in the Vocabulary of the Egyptian Language: Texts and Knowledge in the Culture of Ancient Egypt" (formerly known as "Altägyptisches Wörterbuch") located in the Berlin and Leipzig Academies of Sciences. It contains the largest corpus of Egyptian texts worldwide (ca. 1.4 million text words) and is a crucial tool for linguistic, philological, lexicographical, and cultural research within Egyptology. The encoding software BTS (Berlin Text-encoding System) allows inline annotation and stand-

off-annotation. The range of stand-off annotation options includes commentary of particular parts of texts but also the development of whole annotation layers. The first annotation layer has been developed in Göttingen for the annotation of figurative language, such as metaphors and metonyms, and allows for the annotation of lexical, cognitive and textual dimensions. This presentation introduces the annotation tool for the first time and considers perspectives for the future implementation of the data the tool produces.

10:30-11:00
Coffee break

11:00-11:30
Jonathan Robie, biblicalhumanities.org
XML, Treedown, CSS, and XQuery: Markup and Markdown for creating, visualizing and querying syntax trees

Abstract
This talk explores XML and simpler text-based methods for creating and representing syntax trees, discussing what each approach does best and how they can best be used together for creating, visualizing, and querying treebanks. This talk demonstrates the advantages of XML, presents a language called Treedown designed to simplify display and creation of treebanks, then demonstrates conversions from XML to Treedown and from Treedown to XML.
XML is widely used in digital humanities because of its ability to represent the kinds of hierarchy and sequence that regularly occur in texts, allowing text and metadata to be used together for formatting or querying a text. Just as JSON was designed to represent programming objects, XML was designed to represent the kinds of relationships that occur in text, allowing metadata to describe structure and relationships in the text, provide identifiers, or use additional metadata to record interpretations or associated information. Most syntax trees for Greek are stored as XML, including the Ancient Greek and Latin Treebank (https://perseusdl.github.io/treebank_data/), PROIEL (https://github.com/proiel/proiel-treebank/), Global Bible Initiative syntax trees (), Cascadia Syntax Diagrams, and Lowfat syntax trees. XML technologies provide ways to create, visualize, and query syntax trees using existing, general purpose tools. We demonstrate different kinds of markup, including treebank markup, showing multiple displays of the same data using CSS stylesheets, then show syntactic queries that leverage syntax trees using XQuery and XPath.
Treedown (https://github.com/biblicalhumanities/treedown) is a text-based language originally designed to simplify display of syntax trees while also maintaining sentence order. We demonstrate displays of Treedown, including one that interactively unfolds a sentence at various levels of granularity. When we first invented Treedown, we quickly realized that it could also be used as a “little language” to create syntax trees quickly, providing the same syntactic information as an XML tree. By merging the Treedown representation of a passage with morphological analysis, full-featured XML treebanks can be created, and the resulting XML trees can be queried and displayed using the XML toolchain. We demonstrate stylesheets that

display XML as Treedown and a parser that creates XML from Treedown.

11:30-12:00
Marco Passarotti, Università Cattolica di Milano (Catholic University of Milan)
A Practical introduction to resources and tools for Latin at the CIRCSE Research Centre

Abstract
The talk provides an overview of the activities ongoing at the CIRCSE research centre of Univeristà Cattolica del Sacro Cuore (Milan, Italy). See: http://centridiricerca.unicatt.it/circse_index.html. In particular, the following linguistic resources and tools for natural language processing are presented:
- Word Formation Latin: a Marie Curie IF funded project that enhances the currently available morphological analyser for Latin Lemlat (www.lemlat3.eu) with derivational morphology information;
- the Index Thomisticus Treebank: the theoretical background supporting the syntactic and semantic annotation of the syntactically annotated portion of the opera omnia of Thomas Aquinas;
- IT-VaLex and Latin Vallex: two lexica automatically induced from the Index Thomisticus Treebank. IT-VaLex provides the syntactic subcategorisation patterns of the verbs occurring in the treebank. Latin Vallex is a valency lexicon, reporting the argument frames of the valency-capable lemmas in the treebank (plus excerpts from the Latin Dependency Treebank of Classical Latin).
In the end, the talk gives an insight towards dynamic natural language processing as the most urgent challenge for addressing the problems raised by dealing with a language like Latin, which shows a wide diachronic span.

12:00-13:30
Lunch

13:30-14:00
Dirk Roorda, Koninklijke Nederlandse Akademie van Wetenschappen (Royal Netherlands Academy of Arts and Sciences)
Programming theologians. What they want, do and need

Abstract
Text-Fabric is a datamodel, file format and processing tool for ancient texts and annotations. The approach is radical stand-off markup, and as such it is optimized for slicing and dicing the annotation data and sharing modules of annotations. The stand-off method does not have a good reputation for dealing with changing texts. Yet Text-Fabric is also able to deal with versioning in a graceful way. We demonstrate a new practice of dataprocessing hands-on, based on the Hebrew Bible and a huge set of linguistic annotations.

14:00-14:30
Lydia Müller, Universität Leipzig (Leipzig University)
WebAnno and ASV Toolbox: language independent NLP tools

Abstract
Under-resourced languages, such as historical languages good language models are often not available. However, they are required for most NLP tools to produce reliable annotations or results. For example, POS tagging with language specific of universal POS tag sets requires word to POS tag alongside with probabilities.

A first approximation can be models of a modern form. Afterwards, those annotations can be corrected manually. The curated annotation provides the possibility to train accurate models for historical languages. Training and evaluation of POS tagger models as well as POS tagging using such a model can be conducted with the ASV Toolbox. The produced annotation can be subsequently corrected and curated with WebAnno. Thus, the combination of language independent NLP tools and a system for curating annotations allows a efficient generation of language models for POS information and other annotations.

Apart from annotations, the ASV Toolbox allows also for further task the possibility to train models. As the tagger, they can be used afterwards to conduct the task. Among them is Pretree, a tool to train and use pretrees for baseform reduction and decomposition of compound words, Topic Models, to train and infer topic models on documents and JLanI for language identification.

14:30-15:00
Stefan Schnell, University of Melbourne
GRAID and RefIND: Corpus annotation for cross-linguistic research at the discourse-grammar interface

Brief summary
The author presents a project for a corpus-based typology. The GRAID annotation scheme is detailed to allow language comparison based on concepts rather than language-specific categories.

15:00-15:30
Coffee break

15:30-16:00
Petr Zemánek, Univerzita Karlova (Charles University)
Complex linguistic corpus or a virtual collection: Digital representation of a collection of Assyrian cuneiform tablets

Abstract

The current paper offers a preliminary outline of the principles of a linguistic annotation within a project of a complex database system of a collection of Old Assyrian cuneiform tablets from Kültepe at Charles University. The tablets contain legal documents and correspondence among the members of the so-called Assyrian Trade Network in the 19 th century BCE.
In case of the cuneiform languages, several layers organizing the information are necessary. First of them is the level of individual signs, which are subsequently organized to words and sentences. The basic problem hindering the process is the fact that many tablets are seriously damaged and in case of several genres (including legal documents and correspondence) a reasonable reconstruction is excluded (unlike in narrative genres, where usually several versions of a text are available). On the margins of such lacunae, words or their parts with uncertain roles in the text can be met, in some cases certain degree of reconstruction is possible. The problem of a relatively stable readings and partial or full reconstruction of forms creates a major problem for the annotation of such zones. The degree of uncertainty (present also in case of readable signs) is high and proper definition of linguistic properties of a form is difficult. This holds especially for the syntactic annotation, where the property should be linked to the remaining part of the sentence. Possibilities of the annotation of such parts are discussed.

16:00-17:00
Discussion

July 7 – Raum 702
09:00-9:30
Anke Lüdeling, Carolin Odebrecht, Laura Perlitz, Gohar Schnelle, & Zarah Weiß (Humboldt University Berlin & University of Tübingen)
A Digital infrastructure to support the study of historical German: The RIDGES Herbology Corpus

Brief summary
The authors presents the RIDGES Herbology Corpus, which contains herbal texts written in historical German. The corpus shows an example of how a multi-layered annotation should be designed for a historical language.

9:30-10:00
John Lee, 香港城市大學 (City University of Hong Kong)
Syntactic patterns in classical Chinese poems

Abstract
It is widely believed that different parts of a classical Chinese poem vary in syntactic properties. The middle part is usually parallel, i.e. the two lines in a couplet have similar sentence structure and part of speech; in contrast, the beginning and final parts tend to be non-parallel. Imagistic language, dominated by noun phrases evoking images, is concentrated in the middle; propositional language, with more complex grammatical structures, is more often found at the

end. We present the first quantitative analysis on these linguistic phenomena—syntactic parallelism, imagistic language, and propositional language—on a treebank of selected poems from the Complete Tang Poems. Written during the Tang Dynasty between the 7th and 9th centuries CE, these poems are often considered the pinnacle of classical Chinese poetry. Our analysis affirms the traditional observation that the final couplet is rarely parallel; the middle couplets are more frequently parallel, especially at the phrase rather than the word level. Further, the final couplet more often takes a non-declarative mood, uses function words, and adopts propositional language. In contrast, the beginning and middle couplets employ more content words and tend toward imagistic language.

10:00-10:30
Martin Haspelmath, Max-Planck-Institut für Menschheitsgeschichte, Jena (MPI-SHH) & Leipzig Universität (Leipzig University)
Comparative concepts in cross-linguistic grammatical databases and in glossing

Brief summary
The author shows the difficulties of applying the same linguistic categories to different languages. By means of many examples, the same linguistic labels are showed to capture different concepts in different languages.

10:30-11:00
Coffee break

11:00-11:30
Mattis List, Max-Planck-Institut für Menschheitsgeschichte, Jena (MPI-SHH)
Annotation and analysis of cross-linguistic lexical data in historical linguistics: Towards the establishment of standards and best practices

Abstract
Although historical linguistics has always been a highly data-driven discipline, scholars have been ignoring the importance of standards which hold across subfields and language families. With growing amounts of cross-linguistic digital language data, this becomes more and more evident. As of now, the majority of data produced by historical linguists is only accessible to experts who understand the idiosyncrasies of annotatation practice in specific subfields or specific language families. This hampers both qualitative and quantitative investigations on linguistic diversity. The Cross-Linguistic Data Formats initiative of the Max Planck Institute for the Science of Human History tries to tackle these problems by establishing standards for the representation and analysis of cross-linguistic linguistic data. In contrast to efforts of data collection in the field of natural language processing, our goal is to provide strictly cross-linguistic standards, amenable to all forms of linguistic diversity and not excluding any of the 7000 languages spoken today and in the past. In the talk, we will present the major strategies we use to develop standards and best practices for lexical data, as well as the major

challenges and pitfalls we have to cope with.

11:30-12:00
Dan Zeman, Univerzita Karlova (Charles University)
Universality in space and time – Modern treebanking for ancient languages

Brief summary
The author shows the Universal Dependencies project and its underlying principles. The project is aimed to apply the same morphosyntactic annotation scheme to both modern and ancient languages. Key annotation strategies are emphasis on form rather than meaning, primacy of content words rather than function words, and asymmetrical annotation for coordination.

12:00-13:30
Lunch

13:30-14:00
Volker Gast, Friedrich-Schiller-Universität Jena (Friedrich-Schiller-University Jena)
From temporal annotations to temporal structure: Some explorations

Abstract
In my talk I will propose an annotation scheme for the temporal structure of texts, as well as an implementation of that scheme using GraphAnno (Gast et al. 2016). The annotation scheme is based on Klein's (1994) theory of tense and aspect. I will present two case studies: First, the Timebank corpus (Pustejovsky et al. 2003) was imported into GraphAnno and enriched automatically with information on Topic Time, which allows for a richer and more precise representation of temporal structure. Second, a part of a 19th century novel (W. Collins' "The woman in white") was richly annotated for argument structure and temporal relations. The annotations were then transformed into a two-dimensional structure encompassing narrated time and narrative time. The algorithm transforming the annotations into temporal structures was varied, with the intention to identify the type and amount of temporal information that is needed to correctly represent the temporal structure of a narrative.

14:00-14:30
Justin Cale Johnson, Leiden Universiteit (Leiden University)
Annotating the Babylonian Medical Corpus: Progress and prospects

Abstract
The Babylonian medical corpus represents one of the most complex encodings of the Akkadian language in the entire history of cuneiform, while at the same time presents us with one of the most important groups of texts for understanding the early history of scientific thought. The BabMed Project is currently finishing up a 600 text, 20,000 word online corpus of the medical texts published in BAM volumes 1-6. The first half of these materials are already online at the BabMed website as well as on two infrastructural projects for cuneiform materials, CDLI and

ORACC. The talk focused on both the essential facts of the corpus as well as the ongoing need for access to original image data and Assyriological transliterations, but then went on to ask how higher level annotation could be applied to the Babylonian medical corpus and whether or not annotation frameworks that were used outside of cuneiform studies might be useful for analyzing Babylonian medicine as well.

14:30-15:00
Francesco Mambrini, Deutsches Archäologisches Institut (DAI) & Leipzig Universität (Univerisity of Leipzig)
Trees and Idiolects. Treebank annotation and the study of direct speeches in Ancient Greek literature

15:00-15:30
Coffee break

15:30-16:00
Anton Karl Ingason, Háskóli Íslands (University of Iceland)
Annotating and querying the Icelandic Parsed Historical Corpus and closely related cross-linguistic counterparts

Abstract
This presentation will introduce the annotation and querying infrastructure associated with the Icelandic Parsed Historical Corpus (IcePaHC) project. In addition to the annotation tools, a special consideration wil be given to PaCQL (Parsed Corpus Query Language), a novel query language for carrying out research on parsed historical corpora, an important task for the digital humanities. PaCQL implements and enhances many of the most important features of earlier software that is designed for computational research in historical syntax and combines such functionality with a search engine which employs a fast in-memory index that cuts down waiting time in many realistic research scenarios. A web interface is provided with an automatically created summary of the main quantitative findings, including a visualization and output formats that are suitable for further processing in statistical packages like R and SPSS. The primary goal of this project is to contribute to the development of software tools which are designed from the ground up specifically with the needs of the digital humanities in mind.

16:00-16:30
Neven Jovanović, Sveučilište u Zagrebu (University of Zagreb)
From annotation to learners' corpora

Abstract
It is usually said that 10,000 hours of practice are needed to achieve mastery in a field. How to do this for historical languages, where contact with teachers is necessarily limited? A possible means of support are computer-generated (and assessed) exercises, which will help the student learn, recognize, and produce words, phrases, parts of sentences or even whole sentences,

practicing briefly, but often, and even in situations when they would usually be in "idle speed" (while commuting etc). Such exercises are part of standard learning environments, for example Moodle; in these environments, reporting on user activity is also well supported. The exercise modules, however, seem to expect activities to be created primarily "by hand", to be put together by teachers. Treebanks, vocabulary lists, and similar collections of linguistic annotations offer possibility to create a large number of exercises from authentic (not made-up) language automatically, by retrieving necessary linguistic material from the collections and then transforming it into the format required for import into the learning environment (for example, Moodle Questions XML); the task of the teacher is then simply to select a set of questions for an activity. Such re-use of linguistic annotations will be illustrated on the example of existing Greek and Latin treebanks (PROIEL, Perseus DL, Late Latin Charters Treebank) and word frequency lists (Dickinson College Core Vocabulary). It will be shown as well that, by serving as source for exercises, collections of linguistic annotations easily and naturally connect research and teaching.

16:30-17:30
Final discussion