CLARIN Resources for Classical Latin and Historical German

Brian MacWhinney & John Kowalski
Carnegie Mellon University, Pittsburgh USA
Anke Lüdeling & Uwe Springmann
Humboldt University, Berlin
Detmar Meurers & Zarah Weiss
University of Tübingen

Abstract

The LangBank Project is a collaboration between Carnegie Mellon University, the University of Tübingen, and Humboldt University in Berlin to create web-based corpus resources for the study of Classical Latin and Historical German by both language learners and scholars. These resources are all being made available through the TalkBank CLARIN-B Centre.

1 Introduction

Research and education in the Humanities depends on an understanding of the intellectual roots established by classical cultures, such as Greece, Rome, India, and China, as well as historically important languages such as Latin and Historical German. To understand these cultures, students and scholars need to be able to read texts in the original languages. The LangBank Project seeks to promote students' learning of Classical Latin and Historical German, and to facilitate the ability of more advanced scholars to access a wide range of annotated texts. This new system is creating modern web-based methods for corpus analysis and distribution, online reading support and demand-driven, incidental tutoring of grammar and vocabulary, and learning analytic methods for tracking how students and scholars use the materials.

Work on this project benefits from a variety of existing resources, frameworks, and technologies. Canonicalized texts in Classical Latin are provided by the LatinLit Project, as coded into TEI through the Perseus/PHI Project. Texts in Historical German come from Anke Lüdeling's LAUDATIO Project at Berlin which has incorporated the RIDGES corpus of medieval botanical texts within the ANNIS corpus framework (http://corpus-tools.org) (Zeldes, Lüdeling, Ritz, & Chiarcos, 2009). The current project represents the first step in a more general program for developing similar materials for other historical languages, as well as materials to support the reading and corpus study of modern languages.

2 User groups

We are targeting two major types of users. The first group includes scholars in fields such as history, literature, medicine, and law who need ways of accessing, analyzing, and evaluating detailed material in historical texts. For these users, we rely on the ANNIS system to display alternative levels of linguistic analysis, such as parse trees, morphological structures, coreference – all searchable across corpora in the ANNIS database. Also, for this group, our decoding of the various TEI tags will make it possible to display texts with and without additional information such as alternatives, corrections, and notes. This group of users includes researchers interested in conducting corpus analysis using concordances, frequency counts, and WordSketches – all available through ANNIS.

The second group of users includes learners and their teachers who are interested in locating exercises or examples from original texts, particularly for Latin. Here, the availability of classical texts from Caesar, Plautus, Cicero and others can greatly facilitate work on teaching and learning. For this group, we are developing methods for recording users' vocabulary level and control that can help in preparing them to read new texts.

We are also computing complexity measures that can determine the correct reading level of a given text. Using tagged corpora, we automatically annotate our corpora of Historical German and Classical Latin with over 200 features of linguistic complexity. The features stem from the syntactic, lexical and morphological domain, but also include features of text cohesion and coherence, as well as the modeling of some grammatical aspects highly specific to the given language, such as the ratio of non-finite

clauses in an Early New High German text. The provision of a variety of complexity features as meta information allows users to access texts based on their linguistic properties such as the occurrence of certain grammatical constructions. Calculation of the complexity features for both languages is based on our complexity analysis program (Hancke, Vajjala, & Meurers, 2012).

3 Available Corpora

For Classical Latin, we have produced a full JSON annotation of the LatinLit corpus, as collected by the Perseus Project at https://github.com/PerseusDL/canonical-latinLit. This corpus contains the major works of Classical Latin. These texts were placed into a canonicalized form during the Middle Ages and the digital versions were further annotated across recent decades using 111 different TEI markup codes with widely varying standards and accuracy. To address these problems, we have developed a consistent JSON-based system described below.

For Historical German, all corpora are in the Laudatio repository at http://www.laudatio-repository.org/repository/. They are already in a systematic TEI-based format (Odebrecht, 2014) in the RelAnnis format, and can be converted into all other formats supported by the SaltNPepper framework (https://korpling.german.hu-berlin.de/p/projects/saltnpepper/wiki). The corpora involved include DeutschDiachronDigital, Referenzkorpus Altdeutsch, Referenzkorpus Mittelhochdeutsch, DeutscheDiachroneBaumbank, RIDGES-Herbology, Fürstinnenkorrespondenzkorpus, GerManC, KaJuK, and the Märchenkorpus. We are in continual contact with other initiatives developing historical German corpora and hope to be able to include still more corpora. The current status of our corpus work for Historical German can be examined in the Laudatio repository. We have also incorporated our tagging of the Ridges corpus into ANNIS, but this is not yet viewable online.

4 Normalization

Because neither spelling nor grammar were systematized in Early Modern German, workers in Germany have developed orthographic normalization methods (Bollmann, Dipper, Krasselt, & Petran, 2012; Jurish & Würzner, 2013). Consider the case of a user who wants to find a text on the medical uses of a specific herb (say *absinthium*). Because Historical German spelling is not standardized it is very difficult to find the passages on absinthium. The Ridges herbology corpus, for example, has the following variants for *Wermut* (Modern German for absinthe, wormwood): *wermůt, wermůt, wermut, Wermut, Wermůt, WErmůt, Weremůt, Wermuht, Wermuth, Wer=muth*. These (often unexpected) spelling variations make a reliable manual search impossible. However, if the corpus is annotated with additional normalized forms, vocabulary-based searches are simple and reliable.

Spelling normalization is not a problem for Classical Latin, although there are problems involving canonicalization, capitalization, abbreviation, and sentence segmentation. Once normalized, we subject these texts in both Latin and Historical German to automatic analysis on the lexical, morphological, and syntactic levels. Using the annotations resulting from this analyses we have constructed methods for assisting learners in acquiring the vocabulary needed to comprehend specific texts. We have also begun analysis of these texts in terms of linguistic complexity and readability (Hancke et al., 2012).

5 Sentence segmentation

There are numerous problems with sentence segmentation for Historical German, due to the lack of proper punctuation. These problems make application of automatic parsing inaccurate and unreliable. The introduction of proper punctuation marks for segmentation has proven to be a non-trivial exercise itself, due to the high amount of structural ambiguities, grammatical differences between Early and Late New High German and disagreements in contemporary research on the definition of a sentence. To address this, we have formulated and tested a system of annotation guidelines, which allow users to annotate t-units for Early New High German. In the document formalizing this method, the t-unit definition has been broadened to include solitary phrases, such as interjections. While first tests showed promising results, the next steps include the proper evaluation of inter-annotator agreement. We have also applied these same criteria to texts from Classical Latin. In Latin, segmentation in canonical texts

is generally more reliable and consistent, but some of the same issues arise and can be addressed through use of the same criteria.

6 Creation of a JSON database for Classical Latin

For Classical Latin, our primary focus has been on the creation of a new and fully systematized JSON database derived from the TEI codes of the Perseus/PHI corpus within LatinLit. The shape of this work can be viewed at http://sla.talkbank.org/Latin/About/. TEI describes a set of guidelines specifying methods to encode texts in XML. It has been developed since the 1980s to include over 500 tags. While this rich set of tags provides incredible expressive power to encode a document, it can make it difficult to perform fundamental tasks. For instance, when trying to read or analyze a segment of text, a user may inadvertently include annotating text marked by the <NOTE> tag, or include all variations of a portion of text, as marked by all the children of the <CHOICE> tag. They may pull out too little of the original text, if they ignore the <ADD> tag or too much if they include the <DEL> tag. They may miss punctuation, if they ignore the <QUOTE> tag. Without understanding all the tags in the document, the reader cannot know with confidence if they have the canonical text of a segment. Adding to this challenge, there are numerous ways segments themselves can be defined in TEI, and numerous simple errors in the formatting of tags in LatinLit.

Our goal is to provide easy access to the canonical text of document segments, with the option to get text alternatives and supplementary annotations through a simple document specification. This simplified specification will provide great benefit to a number of communities. In computational linguistics, many applications rely on clean, tokenized texts. For instance, canonical texts of segments allows for testing, training, and improving word alignment tools. Another benefit is to application programmers who may want to create apps with these texts without having to learn and parse the complicated intricacies of a document encoding. For its simplicity and near ubiquitous support by all major programming languages, we choose JSON (Java Script Object Notation) to develop our document spec. We are developing a script to automatically convert all of the TEI documents in LatinLit to our JSON format. The script takes every node in a TEI document that contains text and analyzes the tags along the path to it, extracting information from them. With this script, we will convert the entire Perseus/PHI corpus, supplementing the script tag by tag for each TEI tag we encounter (approximately 111 unique TEI tags in Perseus). The full set of tags found in Perseus/PHI is given at http://sla.talkbank.org/Latin/About/TEI_tags.html

Each JSON document is composed of a *docInfo* object describing the document, followed by the document *content* – a series of *text segment* objects. A given text segment object is composed of (1) the text of the segment, (2) a line ID, (3) an array of TEI objects defining the place of this text segment within books, chapters, books, etc., and (4) a *tok* array of tokens corresponding to each word and punctuation in the segment. Here is an example of a *tok* array for the first two words from Julius Caesar's De Bello Gallico:


```
{
  "txt": "Gallia",
  "type": "word",
  "sp": true,
  "marmot":{"m1":"Ne","m2":"NUMBs|GENDf|CASEn"},
  "tokID": 0
  },
  {
  "txt": "est",
  "type": "word",
  "sp": true,
  "marmot": {"m1":"V-","m2":"PERS3|NUMBs|TENSp|MOODi|VOICa"},
  "tokID": 1
```

},

7 Grammatical Tagging

The *tok* level representations in the JSON database include part-of-speech tags from LatMor (Springmann, Schmid, & Najock, 2016), which have been disambiguated through MarMot (http://cistern.cis.llmu.de/marmot/), as trained on the PROIEL treebank at https://github.com/proiel/proiel-treebank. We are also including short English word-level translations and fuller free translations on the sentence level from the Perseus/PHI resources. We plan to eventually include grammatical relation tags using the CONLL training set from PROIEL.

8 Document Reader

The newly systematized Perseus/PHI texts can now be read through the Document Reader available at http://sla.talkbank.org/Latin/Reader. Using this facility, the learner can read through Caesar's *Commentarii de Bello Gallico* sentence by sentence or paragraph by paragraph. Each word is given a morphological analysis from MarMot and a brief English translation. There is also a sentence-level free English translation. This is an implementation of the Hamiltonian method for learning Latin praised by John Stuart Mill (Blum, 2008). The Hamiltonian method deemphasizes the role of grammar learning and emphasizes the role of reading of a large quantity of texts. This is the method that LangBank is implementing.

In addition to the basic reading tool, we plan to construct personal vocabulary inventories that will help learners work on flashcards for new vocabulary for new materials without having to repeat already known vocabulary words. For grammatical analysis, we can already provide dependency graph analyses through the CLAN programs. We are also planning on adding TTS to allow learners to listen to new texts.

References

Blum, E. (2008). The new old way of learning languages. *American Scholar, 77*, 80-88.

Bollmann, M., Dipper, S., Krasselt, J., & Petran, F. (2012). *Manual and semi-automatic normalization of historical spelling-case studies from Early New High German.* Paper presented at the KONVENS.

Hancke, J., Vajjala, S., & Meurers, D. (2012). *Readability classification for German using Lexical, Syntactic and Morphological features.* Paper presented at the Proceedings of the 24th International Conference on Computational Linguistics (COLING).

Jurish, B., & Würzner, K.-M. (2013). Word and Sentence Tokenization with Hidden Markov Models. *JLCL, 28*(2), 61-83.

Odebrecht, C. (2014). *Modeling linguistic research data for a repository for historical corpora.* Paper presented at the Digital Humanities 2014, Lausanne.

Springmann, U., Schmid, H., & Najock, D. (2016). LatMor: A Latin finite-state morphology encoding vowel quantity. *Open Linguistics*.

Zeldes, A., Lüdeling, A., Ritz, J., & Chiarcos, C. (2009). ANNIS: A search tool for multi-layer annotated corpora.