CHAPTER 1

Historical Languages, Corpora, and Computational Methods

This book explores the challenges of developing the field of Latin Computational Linguistics. My core aim is thus methodological: to show how it is possible to fruitfully combine computational methods with historical language data, particularly synchronic and diachronic data from the Latin language. This is a challenging task and very little work has been done in this direction so far. One of the reasons is that many still see an opposition between the two fields of Computational Linguistics and Latin Linguistics (and Historical Linguistics in general), which would have diverging aims and consequently different methods. For such a hypothetical opposition to hold, Computational Linguistics and Latin Linguistics should be the results of a categorization of the field of linguistics into complementary subfields, just like, say, synchronic vs. diachronic linguistics, Romance linguistics vs. Germanic linguistics, and so on. However, I believe this not to be the case.

Computational Linguistics aims at designing, implementing, and applying computational models for natural languages. This includes developing computer systems that automatically process languages (from a morphological, syntactic, semantic, or pragmatic point of view), as well as building, enriching, and using corpus data (Corpus Linguistics), and much more. Therefore, Computational Linguistics offers a series of methods that can, at least in principle, be applied to a variety of language data. Data from historical languages, however special, do not constitute a conceptual exception to this, and the innovations and adjustments required in such a transition, as well as its limitations, are the principal topics of this book.

1.1 Challenges of Latin Computational Linguistics

Rather than being two separate categories in a one-dimensional classification of linguistics, Computational Linguistics and linguistics of Latin can thus coexist as the result of a bi-dimensional categorization, where the methods chosen (first dimension) are computational and the language studied (second dimension) is Latin. This holds (at least partially) for other historical languages as well.
Why would such a combination be beneficial and why has it not been appropriately investigated before? Why is there not a particular interest in defining, say, Italian Computational Linguistics or Portuguese Computational Linguistics, and why does this book focus on defining Latin Computational Linguistics? The reason is that, when applying computational methods to research on Latin data, a number of questions and challenges arise, which are partially specific to the Latin language, its scholarly community, and the nature of its texts, and partially shared by other historical languages. What are these challenges? In this section I will give a brief overview of the main challenges which will be analytically be dealt with in the rest of the book.

*Linguistic features.* A large part of Computational Linguistics research has been developed for English, or at least tested on this language. Since English is an inflectionally poor language with a so-called fixed word order, many computational models developed for this language tend to pay little attention to the role played by morphology in a number of linguistic phenomena. Instead, if we try to build a computational resource or model for Latin, the morphological layer of annotation is an essential one. For this reason, it is important to resort to annotation schemas developed for other morphologically rich languages. An example is Czech, for which annotation schemas and treebanks have been created. The decision to follow existing standards developed for Czech was taken by the creators of the three Latin syntactically annotated corpora which will be presented in section 2.3.3.1. In a similar vein, when I designed the corpus-based valency lexicon presented in chapter 3, I recorded linguistic features particularly relevant for further investigations on Latin verbs; for example, the lexicon records the order of verbal arguments and their morphological features through specific patterns. These are important examples of some of the challenges researchers face when building computational resources for Latin.

*Research questions, commercial applications, and academic tradition.* Computational Linguistics has mostly focussed on current languages. The interest in developing computational resources and methods for the most widespread extant languages is motivated, among other things, by their applications to various areas of society with important commercial uses, such as Language Technology: computational lexicography, machine translation, speech recognition, virtual assistants, and text summarization, to name just a few examples. On the other hand Historical Linguistics has not fully explored computational approaches. The business picture for extinct languages is different from the one for current languages due to the lack of direct commercial interests. Furthermore, the nature of the linguistic investigation for historical languages with