A Corpus-Based Foray into Latin Preverbs

6.1 New Paradigm

After presenting a computational lexicon for Latin verbal valency in chapter 3 and a system for learning SPs from Latin treebanks in chapters 4 and 5, this chapter gives an example of a quantitative investigation that uses the lexicon data. Given the extremely limited number of quantitative corpus studies concerning linguistic phenomena in Latin, the principal aim of this chapter is to show the potential of such an approach. Therefore, rather than as an exhaustive interpretation of the diachronic development of Latin prefixed verbs, the present study should be considered for its methodological contribution, which shows what can be achieved by using new source data (treebanks and the valency lexicon) and new techniques (multivariate statistics). This methodological focus also explains the emphasis which I intentionally placed on the research design rather than on the steps of a traditional analysis. This case study is meant to show an instance of the collaborative paradigm illustrated in section 1.4.

We can imagine the research process as a circle containing the following phases:

0. Create corpora, annotate them
1. observe the phenomenon, assign the label (e.g. ‘preverb’)
2. do preliminary theoretical work, identifying the phenomenon in different languages and constructing theories on it
3. gather data manually (examples)
4. interpret the example data in light of the theories
5. formulate theoretical hypotheses
6. process corpus data
7. carry out quantitative studies on corpus data
8. prepare first interpretation of the results
9. prepare further interpretation of the results and theoretical considerations such as: “is the label ‘preverb’ adequate?” Back to step 1.

I consider step number 0 to be independent from the particular phenomenon under investigation, and therefore this step does not enter the iterative process.
Given the methodological aims of the book, this study does not cover the whole process; instead, it focusses on phases 1–2 and 5 to 8. It does not include a complete comparison between the results offered by the analyses and the statements contained in previous studies, although it does mention some examples. Moreover, it illustrates the quantitative methodology necessary for further studies on this phenomenon and offers a basis for reconsidering the theories in the light of the analysed corpus data.

I define preverbs as prefixes of verbal stems. For example, \textit{ab-} ‘from’ is prefixed to the verb \textit{eo} ‘go’ in the prefixed verb (PV) \textit{abeo} ‘go from, depart’. How do these verbs express spatial complements? Does \textit{abeo} express the place from which the subject departs with a simple ablative, or with a PP introduced by \textit{ab} or another preposition? To what extent does the preverb affect the argument realization of the PV? In this chapter I will show how a quantitative analysis of corpus data can contribute to the field of Latin Computational Linguistics by addressing these questions as examples. The investigation presented here will testify the challenge of applying quantitative Corpus Linguistics to Latin data. In addition, this work will employ appropriate statistical techniques that measure the degree of uncertainty in the data, in order to overcome the issues related to small diachronic corpora.

One way to explore this topic is by making some theoretically grounded claims and then illustrate text examples, which is the usual practice in so-called qualitative studies. Another way is to refer to a corpus to carry out a systematic, quantitative account of the arguments of PVs, which means that the corpus is the starting point of an analysis that aims at quantitatively describing the data, and then generalizing the results to the whole population. This is achieved by using statistical techniques that help detect significant patterns in the data. In keeping with the general aim of this book, I will choose the latter approach and use the corpus-based lexicon I described in chapter 3 as the basis for the investigation. The quantitative approach allows me to highlight the advantages of using statistical techniques to explore a complex phenomenon, and quantify the role played in it by several factors.

That the phenomenon of preverbation is complex is made clear by the definition of preverbs given in Booij and van Kemenade (2003): “morphemes that appear in front of a verb, and which form a close semantic unit with that verb”. This definition makes an explicit link between the morphological status of preverbs and their semantic properties with respect to the verb they are prefixed to. Although preverbation traditionally pertains to word formation and is therefore a morphological phenomenon, PVs offer the opportunity to explore interesting aspects of the syntax-semantics and syntax-lexicon interfaces. I will show how relevant interactions between the morpho-syntactic realizations of