CHAPTER 2

Computational Resources and Tools for Latin

2.1 The Role of Latin

If we go back to the origins of Computational Linguistics and Corpus Linguistics, we find that, among all languages, Latin played a truly crucial role. Latin is in fact the language of the first electronic corpus, the *Index Thomisticus* (IT) (Busa, 1980), whose development started in the late 1940s and continued for half a century. The IT contains Thomas Aquinas’ *opera omnia*, or 11 million word tokens, a huge figure for those times and still a considerable one for our times. It was first encoded in punched cards, then on magnetic tapes, and finally in digital format.

In spite of this important precedent, after an early phase computational and corpus research has been focusing on modern languages, especially English. Researchers have worked on both building larger and larger corpora, and developing Natural Language Processing (NLP) tools for automatically annotating them at all levels, from lemmatization and morphological analysis to syntactic, semantic, and pragmatic analysis. Research in developing computational resources and tools for Latin for a long time has not kept up with the advances in the field of Computational Linguistics for modern languages. Several different projects aiming at building computational resources and tools for Latin have only started quite recently, and this chapter aims at giving an overview of their achievements.

2.1.1 Why Are Computational Resources Needed?

Krauwer (1998) introduces the idea of a *Basic LAnguage Resource Kit* (BLARK), which he defines as “the minimal set of language resources that is necessary to do any precompetitive research and education at all” (Krauwer, 2003, 4). Such language resources range from spoken and written language corpora, monolingual and bilingual dictionaries, morphological analyzers, parsers, annotation standards and tools for corpus exploration and exploitation. Different languages require different resources and some resources may only be relevant for certain languages.

Ancient languages do not have complete BLARKs. The reasons for this lie in the cost (and challenges) of creating historical corpora and in the lack of investments to implement new and large computational resources for these languages. As Passarotti (2010) observes, of all ancient dead languages, Latin
is definitely in a lucky position, because its BLARK is in fieri and relatively rich, as the overview in the next sections attests to.

2.2 Corpora and Digital Editions for Latin

The first step towards developing computational resources for Latin consists of collecting the textual material in a format that can be handled by NLP tools. For this, it is necessary to digitize part of the immense amount of Latin texts produced throughout the long history of this language. Those texts were not originally produced digitally, which means that a conversion from paper to electronic format is necessary. This labour-intensive task typically involves manual transcription or Optical Character Recognition (OCR); see Piotrowski (2012, 25–52) for a comprehensive overview of the issues and the techniques used to digitize historical texts. Here, I will briefly review some of the digital editions of Latin works available today.

The first example I will consider in this brief overview is the *Index Thomisticus* (Busa, 1980). It is available online¹ and contains the best critical editions of the *opera omnia* by Thomas Aquinas (118 texts), together with works by other authors related to Thomas (61 texts), for a total of around 11 million words. The online version can be searched by word forms.

Some of Latin digital editions were conceived to be used by philologists and therefore contain rich information on the tradition of the texts. For example, the projects *Musisque deoque* and *Poeti d'Italia in lingua latina tra medioevo e rinascimento*,² coordinated by scholars from the Italian universities of Padova, Trieste, Venezia Ca’ Foscari, and Verona, collect the digital editions of Latin works by Latin authors from the Classical era and works by Italian authors between Dante Alighieri’s birth (1265) and the first half of the sixteenth century. For each work the critical text is provided, including the codicological variants and, if possible, scholia and ancient glosses, as well as the Italian translation. On the website it is possible to search for the passages containing certain words (and all the variants obtained via inflection, compounding and derivation); it is also possible to impose constraints on the search such as the metre or the position in the line.

¹ http://www.corpusthomisticum.org/.
² http://www.mqdq.it/mqdq/home.jsp.