Barbara McGillivray / Adam Kilgarriff

Tools for historical corpus research, and a corpus of Latin

We present LatinISE, a Latin corpus for the Sketch Engine. LatinISE consists of Latin works comprising a total of 13 million words, covering the time span from the 2nd century B. C. to the 21st century A. D. LatinISE is provided with rich metadata mark-up, including author, title, genre, era, date and century, as well as book, section, paragraph and line of verses. We have automatically annotated LatinISE with lemma and part-of-speech information. The annotation enables the users to search the corpus with a number of criteria, ranging from lemma, part-of-speech, context, to subcorpora defined chronologically or by genre.
We also illustrate word sketches, one-page summaries of a word's corpus-based collocational behaviour. Our future plan is to produce word sketches for Latin words by adding richer morphological and syntactic annotation to the corpus.

1. Introduction

Latin is the language of the first electronic corpus, the *Index Thomisticus*, compiled by Father Roberto Busa between the late 1940s and the 1970s, and typically considered to have marked the beginning of linguistic computing (Lüdeling and Zeldes 2007). Since those times, corpus linguistics and computational linguistics have developed into mature disciplines, and a number of modern languages have been provided with large annotated corpora and computational tools. In particular, sophisticated corpus query systems have been created that allow linguists to carry out advanced searches on corpora. Despite its promising start, Latin, like other dead languages, has partially been left behind in this process, especially for what concerns the availability of large corpora with rich syntactic and semantic information and of advanced corpus query systems.
This paper focuses on the Sketch Engine, a leading corpus query tool that is widely used in a number of lexicographic projects and corpus research (Kilgarriff et al. 2004). We present the functions and the resources we have added to the Sketch Engine to meet the needs of historical corpus research. In particular, we illustrate LatinISE, the first historical corpus included in the Sketch Engine. LatinISE is a 13-million word Latin corpus whose texts range from the 2nd century B.C. to the beginning of the 21st century A.D. LatinISE has been automatically annotated with state-of-the-art Natural

Language Processing (NLP) tools: a lemmatizer and a part-of-speech (POS) tagger.
The rest of the paper is organized as follows: section 2 gives the background on existing corpora for Latin and motivates the project for LatinISE; section 3 illustrates the Sketch Engine and its main functionalities; section 4 describes how we have collected and automatically annotated LatinISE and, finally, section 5 concludes by summarizing future research.

2. Latin corpora

Latin can be considered a less-resourced language from a computational point of view if we compare it with modern languages like English. However, for a dead language, the range of available corpora for Latin is quite large.
Over the past years several projects have dealt with digitizing the immense amount of texts produced throughout the history of the Latin language. These projects have created a large number of digital editions, which can be browsed and searched through *ad hoc* search engines: thanks to such tools, philologists, linguists and literary scholars can look up occurrences of single word forms or sequences of word forms to extract their contexts in the texts (concordances).
Some Latin digital editions have been designed for philologists and therefore contain rich information on the tradition of the texts. An example is *Musisque deoque* (http://www.mqdq.it), collecting Latin works by poets from the archaic to the modern era. Other projects have chosen a particular edition rather then displaying the complete philological information. An example is the *Library of Latin Texts* (CTLO 2010), containing more than 50 million words. This corpus is available on CD-ROM and is searchable by lexical forms and chronological eras. Another private collection of Latin texts is the *Bibliotheca Teubneriana Latina* (CETEDOC 1999), containing 10 million words and published as a CD-Rom. Among the open-access digital collections it is worth mentioning the *Perseus Digital Library* (Bamman and Crane 2008; http://www.perseus.tufts.edu), consisting of 10 million words. Thanks to the morphological analyzer Morpheus, the Perseus Digital Library is searchable by word forms and lemmas. Around 53,000 words belonging to several classical works collected in the Perseus Digital Library have been morphologically and syntactically annotated in the Latin Dependency Treebank (Bamman and Crane 2006).
The *Index Thomisticus*, consisting of 11 million lemmatized words and collecting Thomas Aquinas' *opera omnia*, is still among the largest existing Latin corpora, and it is now available online (Busa 1974-1980; www.corpusthomisticum.org). A morphologically and syntactically

annotated portion of this corpus is available as the *Index Thomisticus* treebank (Passarotti 2007; http://itreebank.marginalia.it).
The third treebank for Latin was created as part of the PROIEL (Pragmatic Resources of Old Indo-European Languages) project (http://www.hf.uio.no/ifikk/english/research/projects/proiel/) and contains around 90,000 words mainly from the Jerome's translation of the Bible (Haug and Jøndal, 2008).
As attested by this brief and non-exhaustive overview, a considerable number of Latin corpora are available to linguists nowadays. However, the searches that are possible on them are limited by their annotation. The majority of these collections contain raw text, so the search options are limited to word forms; in some cases the user can look for lemmas, while more advanced searches are only possible in the three treebanks, which are very limited in size (between 53,000 and 120,000 tokens).
The aim of our project was precisely to fill this gap and build a large, richly annotated corpus of Latin covering an extensive time span. To do that, we decided to apply state-of-the-art NLP tools, which allow fast and consistent automatic annotation. In addition, we wanted to make the corpus searchable through a flexible and sophisticated corpus query tool: the Sketch Engine.

3. Sketch Engine

Since its start in 2003, the Sketch Engine has been in use in several dictionary projects and its value for lexicography is illustrated in Kilgarriff and Tugwell (2001), among others. One of the advantages of the Sketch Engine is that it is provided with a wide range of corpora, and is able to handle large amounts of data (the largest corpus to date contains 8 billion words). The web interface allows the user to upload her own corpus or to automatically build it from the web. In addition, the Sketch Engine provides highly developed search options on the corpora, which makes it an ideal tool for dictionary making. These options include word form, lemma, phrase and CQL (Contextual Query Language) search, as well as filters on contexts of a target word, such as the size of the left/right context, the lemmas and parts-of-speech of the words in the context. The output of such searches is a set of concordances, with customizable view and sorting settings.
In addition to these advanced concordance features, the Sketch Engine provides *word sketches*, its distinctive feature. Word sketches are one-page automatic corpus-based accounts of a word's grammatical and collocational behaviour.

[FILE
word_sketch_bw.tiff]

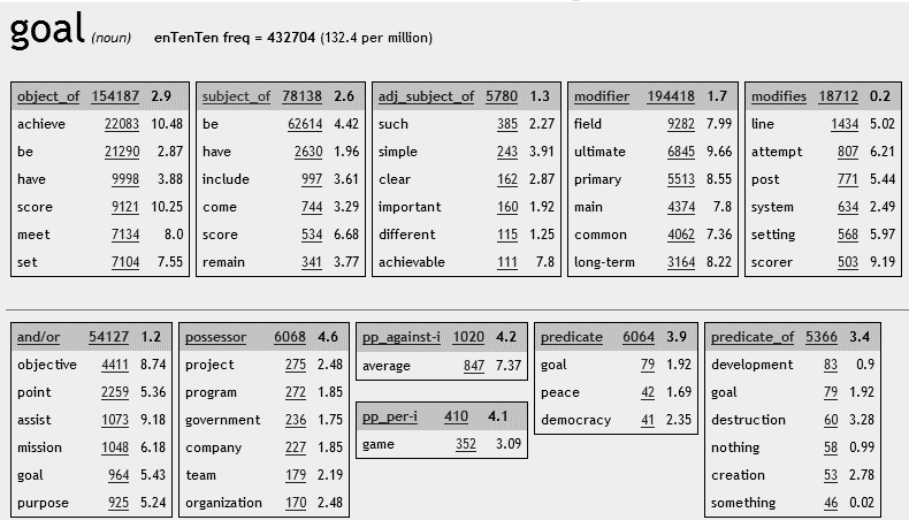

Figure 1. Example of word sketch for the noun *goal* in the enTenTen corpus.

Figure 1 shows the word sketch for the noun *goal* in the enTenTen corpus for English, containing over 3 billion tokens. The word sketch is organized by grammatical relation and has been produced from a syntactically parsed corpus. Each section of the word sketch shows which words stand in a particular grammatical relation with the target word *goal*. For example, the section "object_of" contains verbs whose syntactic object in the corpus is *goal*. Each such collocate is shown with its corpus frequency and salience.
This example gives an idea of the potentialities of word sketches for corpus-based linguistic studies on words' behaviour. Along the same lines as word sketches, *sketch differences* show the differences in the corpus behaviour of two target words, for example by highlighting which collocates are shared by the two words and which ones are specific to only one of them.
No large historical corpus has been provided with such a rich range of search options so far, and our project aimed at making Latin the first dead language to be included in the Sketch Engine.

4. LatinISE: a Latin corpus in the Sketch Engine

In this section we will go through the project phases, from explaining how we collected the texts (section 4.1), to describing its metadata and subcorpora (section 4.2), illustrating the morphological annotation (section

4.3) and POS tagging (section 4.4), and finally exemplifying how the corpus can be searched and displayed (section 4.5).

4.1 Collecting the texts

The first phase of our project consisted in the collection of the texts. These were assembled from three online digital libraries: LacusCurtius (http://penelope.uchicago.edu/Thayer/I/Roman/home.html, by Bill Thayer), IntraText (http://www.intratext.com), and Musique Deoque (http://www.mqdq.it). These digital libraries contain texts from standard editions and cover a wide time span, as well as a variety of genres. In this respect, they were ideal for our purposes of creating a large and wide-ranging corpus for Latin.

The texts had to be converted from HTML format into the verticalized format required by the Sketch Engine. While converting the HTML files, special care was devoted to keeping the metadata mark-up specifying authors, title, books, sections, paragraphs and lines (for poetry). In the verticalized text each line corresponds to a token, a punctuation mark or a tag, and looks like this:

```
<character name="Th">
<line>
praemia
si
cessant
<g/>
,
```

The <g/> tag always precedes punctuation marks and has the effect of suppressing space characters between two tokens.

4.2 Metadata and subcorpora

In a historical corpus, especially a diachronic one, rich metadata annotation is essential, given the specifically literary and/or diachronic interest of the users. All three digital libraries provide the texts with metadata information, which was thus extremely helpful. The metadata were also used to automatically eliminate duplicates of the same texts, an important task in automatic corpus building.

Our metadata cover author, title of the work, genre (prose or poetry), era, date of the work (when available), and century. The oldest text in our corpus are the *Senatus consulta de Baccanalibus* (186 B. C.), and the most recent one is *Dominus Iesus* (2000), by the Vatican Congregation for the Doctrine of the Faith. Below we show an example of how the metadata information is encoded in the corpus for the first text from LacusCurtius (LC):

<doc id="LC" n=1 author="uncertain" title="Einsiedeln Eclogues" genre="poetry" era="Romana, Postclassica" date="cent. 1 A. D." century="cent. 1 A. D.">

Our classification in eras follows the one adopted in IntraText and include *Romana Antiqua* (VII-II cent. B. C.), *Romana Classica* (I cent. B. C.), *Romana Postclassica* (I-VI cent. A. D.), *Mediaevalis* (VII-XIV cent. AD), and *Nova* (XV-XXI cent. AD).
The Sketch Engine allows the corpus builder to define subcorpora according to specific metadata features. For example, the prose subcorpus has 9,935,401 tokens[1], while the poetry subcorpus has 3,818,603 tokens.

4.3 Morphological annotation

In order to annotate the corpus, we used state-of-the-art NLP tools. Unlike manual annotation, automatic methods have lower accuracy but are far faster and cheaper. Also, automatic annotation can be easily updated as the input corpus increases or changes.
We aimed at enriching the texts with lemmas and POS tags. For the lemmatization phase, we used the PROIEL Project's morphological analyser developed by Dag Haug's team; for those word forms that were not recognized by this analyser, we used Quick Latin (http://www.quicklatin.com/). The input to the analysers was the verticalized text; for example the output of the phrase *sumant exordia fasces* 'let the fasces open the year' looked like this:

> sumant
> sumo<verb><3><pl><present><subjunctive><active>
> exordia
> exordium<noun><n><pl><acc>
> exordium<noun><n><pl><nom>
> exordium<noun><n><pl><voc>
> fasces
> no result for fasces

For each word form the analyser gave all possible analyses, with lemma and POS, as well as other morphological tags (gender, number, case, mood, person and voice).

4.4 POS tagging

Once the possible analyses of each token were available, the next question was how to disambiguate these analyses to find the right one. In particular,

[1] In the Sketch Engine a *token* is a word or a punctuation mark.

we focussed on obtaining the most likely lemma and POS for each token in context. To do this we adopted a machine-learning approach.
Machine-learning POS taggers work on the assumption that if we train a model on some annotated text (training set), it will learn patterns of regularities and will thus be able to tag unseen text.
Lemmatized and morpho-syntactically annotated data for a total of over 242,000 tokens are available from the three Latin treebanks we introduced in section 2: the *Index Thomisticus* Treebank, the Latin Dependency Treebank and the PROIEL Project's Latin treebank. Therefore, we opted to use those data to train TreeTagger (Schmid, 1995), a language-independent POS tagger developed by Helmut Schmid at the University of Stuttgart.
The input to TreeTagger was the output from the morphological analyser, with lemma and POS. Based on the contexts each token occurred in, TreeTagger learned what POS was the most likely among all possible ones. We then assigned the token to that POS and its corresponding lemma.
The output of the annotation was added to the verticalized text so that the first column contained the word form, the second one its POS, and the third one its lemma. For example, the sentence *praemia si cessant,*'if the prizes are lacking, the confidence of skill is dumb', uttered by the character Thamyra in the Einsiedeln Eclogues (1st century A. D.), is represented in the corpus as follows (ADJ is for adjectives, C conjunctions, N nouns, V verbs):

```
<character name="Th">
<line>
praemia    N       praemium
si         C       si
cessant    V       cesso
<g/>
,
```

4.5 Searching LatinISE

The annotation provided in LatinISE allows the user to search for a lemma by its POS. For example, the Latin word *cum* can be a preposition ('with') or a conjunction ('when', 'because'); the user can choose to restrict the search to one POS or to view both the lemma and the POS ('C' and 'PRE') in the concordances. In the latter case the output would look like Figure 2.

[FILE
cum_bw.tiff]

Figure2. Concordances for *cum* ‘with; when, because’ in LatinISE.

A wide range of possibilities are offered by the view options, where the user can display different metadata information (title of the work in Figure 2), the size of the context, the order of the concordance lines by context, and so on. In addition to simple search on word forms, lemmas, and phrases, it is possible to specify the left/right context of a word by the lemma, POS and number of tokens in its context. This allows the user to extract syntactic constructions like *dico*/*puto*/*credo* ‘believe, think’+*quod* ‘that’ (Figure 3), and get an overview of the distribution of these constructions in the corpus.

[FILE
dico_bw.tiff]

Figure2. Context-dependent concordance search for the conjunction *quod* ‘that’ followed by forms of the verbs *dico*, *puto* or *credo* ‘think, believe’.

5. Conclusion and future research

We have presented LatinISE, a 13-million token corpus for Latin. LatinISE is the first historical corpus included in the Sketch Engine and was automatically lemmatized and POS tagged using state-of-the-art NLP tools. The texts contained in LatinISE cover a time span of 22 centuries, from Early Latin to the beginning of our century. Its rich metadata and linguistic annotation makes it possible to carry out diachronic studies on various aspects of the Latin lexicon.

We are planning to enrich the annotation with morphological tags (case, number, gender, mood, voice, person) and, ultimately, syntactic relations. This would allow us to produce word sketches, showing the collocational behaviour of Latin lemmas over time.

References

Bamman, David/Crane, Gregory (2006), The Design and Use of a Latin Dependency Treebank. In: Proceedings of the Fifth International Workshop on Treebanks and Linguistic Theories. Prague: Institute of Formal and Applied Linguistics, 67-78.

Busa, Roberto (1974-1980). Index Thomisticus. Stuttgart-Bad Cannstatt: Frommann-Holzboog.

Crane, Gregory/Chavez, Robert F. (2001), Drudgery and deep thought: Designing Digital Libraries for the Humanities. In: Communications of the ACM, 44(5): 34–40.

Centre Traditio Litterarum Occidentalium (CTLO) (2010), Library of Latin Texts CLCLT-6. Turnhout: Brepols.

CETEDOC (1999): Bibliotecha Teubneriana Latina.Turnhout: Brepols.

Lüdeling, Anke/Zeldes, Amir (2007), Three views on corpora: corpus linguistics, literary computing, and computational linguistics. In: Jahrbuch für Computerphilologie, 9 (2007): 149-178 (http://computerphilologie.tu-darmstadt.de/jg07/luedzeldes.html).

Haug, Dag Trygve Truslew/Jøndal, Marius Larsen (2008), Creating a parallel treebank of the old Indo-European Bible translations. In: Calzolari, Nicoletta/Choukri, Khalid/Maegaard, Bente/Mariani, Joseph/Odjik, Jan/Piperidis, Stelios/Tapias, Daniel, (eds.): Proceedings of Language Technologies for Cultural Heritage Workshop (LREC 2008). European Language Resources Association, 27-34.

Kilgarriff, Adam/Rychly, Pavel/Smrz, Pavel/Tugwell, David (2004), The Sketch Engine. In: Williams, Geoffrey/Vessier, Sandra, (eds.): Proceedings of the eleventh Euralex International Congress. Lorient: Université de Bretagne-Sud, 105-116.

Kilgarriff, Adam/Tugwell, David (2001), WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography. In: Proceedings of the ACL workshop on COLLOCATION: Computational Extraction, Analysis and Exploitation. Toulouse, 32-38.

Passarotti, Marco (2007), Verso il Lessico Tomistico Biculturale. La treebank dell'*Index Thomisticus*. In: Petrilli, Raffaella/Femia, Diego, (eds.): Il filo del discorso. Intrecci testuali, articolazioni linguistiche, composizioni logiche. Atti del XIII Congresso Nazionale della Società di Filosofia del Linguaggio. Roma: Aracne, 187-205.
Schmid, Helmut (1995), Improvements in Part-of-Speech Tagging with an Application to German. In: Proceedings of the ACL SIGDAT-Workshop, 47-50.