A System for Latin Document Analysis and Translation

Antoine Bossard¹,² and Keiichi Kaneko²

¹ Department of Information Sciences, Faculty of Science, Kanagawa University, Hiratsuka City, Kanagawa 259-1293, Japan
² Graduate School of Engineering, Tokyo University of Agriculture and Technology, Koganei City, Tokyo 184-8588, Japan
³ To whom correspondence should be addressed. E-mail: abossard@kanagawa-u.ac.jp

Abstract: A long time before English, Latin was the international language of the occidental world. Since Latin is at the root of numerous languages, knowledge of it is of marked benefit when learning foreign languages. Especially because of complex grammar, Latin classes in secondary education remain unpopular. In this research note, we summarise our proposal of a system for automatic Latin document analysis and grammar-aware translation support. In addition to such natural language processing aspects, the proposed system features a very accessible user interface that makes the system suitable for pedagogical purposes.

Keywords: language, system, educational, grammar, inflection, declension, conjugation

Introduction

Even if partly concealed, Latin is ubiquitous within occidental cultures, and more precisely, occidental languages. For instance, the etymology of numerous French and English words has Latin roots. Words such as "audio" (from audire "to listen"), "fluctuation" (from fluctuare "to float") and "domestic" (from domus "house") are some examples.

Yet, Latin has a complex grammar, involving for instance cases (declensions), genders and extensive conjugation (moods, tenses and the numerous corresponding inflections). There are two major consequences to this fact:

1. The knowledge of Latin is difficult to acquire, and as such, the number of Latin students has been steadily decreasing, for instance with France almost shutting down Latin classes ¹⁷.

2. The support of Latin by information technologies is minimal, if not inexistent. For instance, translation engines such as Google Translate and Microsoft Translator do not include Latin in the list of their supported languages. Existing online Latin dictionaries are almost exclusively word-based dictionaries without inflection support ². Exceptionally, the Olivetti electronic dictionary ³ does provide inflection support but remains word-based.

The above two issues are indeed not completely unrelated. It is reasonable to assume that a better technology support with respect to Latin would definitely not harm its popularity, quite the contrary. Hence, the aim of this research project is to provide a system realising the automatic analysis of Latin documents, in fine providing translation support and more generally pedagogical support.

Method: System Overview

In this section, we give a brief overview of the proposed system. The system is described in more detail in our paper ⁴. We have built it mostly by using the functional language Racket and its development environment and libraries ⁵.

First, our system relies on a database of Latin words. This database has been assembled from the list of the 1,000 most frequent Latin words as given by C. Francese ⁶. Importantly, our database has been automatically generated, with each entry being fully expanded: noun forms are not abbreviated, and verbs are given with their full five forms (or four for verbs without supine). For instance, the original verb entry "voco, -are" is expanded to the full 5-form "voco, vocas, vocare, vocavi, vocatum". Each database entry includes grammatical information
such as gender, as well as English semantical information. The database creation process is detailed in Fig. 1: the original XML database (word list) is transformed into a Racket word list with an XSL stylesheet, the obtained Racket word list being then normalised (entry expansion as explained above) by a Racket routine. In addition, a secondary database is used to match irregular inflections with the corresponding primary database entry.

The supported grammar features are listed below.
- Conjugation: indicative, subjunctive, infinitive, imperative, participle, gerund moods, present, future, imperfect, perfect, pluperfect, future perfect tenses, and the active and passive voices.
- Lexical categories: nouns, verbs, adjectives (including comparatives and superlatives), pronouns, prepositions, adverbs and conjunctions.

Most notably, the adjective declension system is the most difficult given the numerous patterns. In fact, one adjective usually induces 6 cases × 3 genders × 2 numbers = 36 forms.

The system input consists in a Latin sentence (or simple words), its grammatical analysis being the output. The input text is pre-processed by a parser to distinguish words, removing punctuation symbols, white space, and normalising capitalization.

Last but not least, the proposed system features a user-friendly online interface, which relies on the server component bundled with the software. For that, we have used the suitable Racket libraries such as web-server/servlet. The system is thus very easy to deploy as it includes a web-server component, and to use as a regular web browser suffices. An illustration of the system interface with a sample output is given in Fig. 2. In this figure, the input Latin sentence is Domus patris nostri. “Our father’s house.” The three words are successfully matched by the system, which provides lexical categories (e.g., “feminine noun” for “domus”), match details (e.g., “singular genitive” for “patris”) and English translation information (e.g., “we” and “our” for “nostri”).

### Experimental Results
We have realised with the proposed system the automatic analysis of the first twenty parts of Cicero’s “Laelius de amicitia,” this corpus consisting of 2,008 words. In total, 64% of the corpus (i.e., 64% of the words) were successfully analysed by the system, that is, successfully matching these Latin words with our database, and thus being able to provide grammatical information supporting text comprehension and translation elements as well. An excerpt of the obtained results is given in Table 1, focusing on the three grammatical categories: noun, verb and adjective. In total, 724 words were not matched by the analysis, thus providing no information, for instance, to the student.

The results for the noun grammatical category are further detailed in Fig. 3. Precisely, for each grammatical case, the number of noun occurrences are

<table>
<thead>
<tr>
<th>Word type</th>
<th>Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun</td>
<td></td>
</tr>
<tr>
<td>masculine</td>
<td>77</td>
</tr>
<tr>
<td>feminine</td>
<td>97</td>
</tr>
<tr>
<td>neuter</td>
<td>43</td>
</tr>
<tr>
<td><strong>Total (noun)</strong></td>
<td><strong>215</strong></td>
</tr>
<tr>
<td>Verb</td>
<td>261</td>
</tr>
<tr>
<td>Adjective</td>
<td>185</td>
</tr>
</tbody>
</table>
given for the three genders: masculine, feminine and neuter.

**Discussion**

Out of the 724 unmatched words, many are duplicates. Precisely, a total of 477 distinct words remained unmatched by the system. In addition, it should be noted that several matched words belonged to more than one grammatical category. For instance, the word “vitas” can be analysed both as a noun (from *vita, vitae* “life”) or as a verb (from *vito, vitas, vitare* “to avoid”). Also, out of the 215 detected nouns, two occurrences of “viri” were found: this noun being matched as both a masculine and feminine noun, it explains that the sum of the three gender totals gives 217 instead of 215. This behaviour can also be seen in Fig. 3. Finally, the analysis also exposed a rather uniform repartition of the various grammatical categories (noun, adjective, pronoun, etc.), with respect to the matched words, obviously.

Compared to the Latin analysis system of William Whitaker which weighs 1.2MB (and with no user interface) against a few kilobytes (90KB) in our case, the obtained results are significantly positive. In other words, our system was able to provide supporting information for 64% of the Latin words of the analysed text at a fraction of the complexity of a previous system.

Regarding future works, the authors plan the following developments:

1. Expansion of the supported grammatical categories, with for instance additional conjunctions and prepositions;
2. Detection of the grammatical case of the word that directly follows a particular preposition in order to provide even more helpful translation information;
3. Experimentation with a class of Latin students so as to quantitatively evaluate the pedagogical impact of the proposed system.

**Acknowledgements**

This research has been partly supported by a grant from the Research Institute for Integrated Science, Kanagawa University (RIIS201706).

**References**