Implementation of a Latin Grammar in Grammatical Framework

Herbert Lange

Department of Computer Science and Engineering
University of Gothenburg and Chalmers University of Technology
herbert.lange@cse.gu.se

1. Introduction

Here we present work in developing a computerised grammar for the Latin language. It demonstrates principles for developing a grammar for a natural language in a modern grammar framework. The grammar presented here can be used as a base for other natural language processing applications in different fields like language learning and in language technology for cultural heritage.

The research idea behind this work was to test to what extend it is possible to transfer the information given in a ordinary school grammar book into the computer-readable form of a grammar in the Grammatical Framework.

2. Grammatical Framework

The Grammatical Framework (GF) is a specialised software system for the development of grammars as well as parsing and translation. It is developed as free and open source software at the University of Gothenburg and provides a grammar formalism roughly in the style of modern functional programming languages like Haskell. A more detailed description can be found in the book by Ranta (2011).

GF uses a flavour of context-free grammars extended by some features, most importantly tables and records. Adding these increases the expressivity so that it is equivalent to Parallel Multiple Context-Free Grammars (PMCFG), ie its expressivity is mildly context-sensitive but still has polynomial parsing complexity (Ljunglöf, 2004, p. 57).

Tables can be used for parametric features like the noun cases while additional record fields can be used for inherent grammatical features like the noun gender. Parametric features in lexical items usually give rise to inflection tables.

Another feature of this formalism is the distinction between so-called abstract and concrete syntax. While the abstract syntax only specifies the rules and their parameters on an abstract level, the concrete syntax gives it a concrete form and specifies how the parts are realised on the surface.

Multiple concrete syntaxes can share the same abstract syntax so that the abstract syntax tree can be used as an intermediate representation to translate between the different languages. The most extensive abstract syntax is the one defined by the Resource Grammar Library (RGL) distributed with GF (Ranta, 2009) of which this Latin grammar is part of.

3. Lexicon

The RGL contains a small reference lexicon that is shared among all its languages. This lexicon contains ca 350 entries including several modern concepts such as "refrigerator". To find plausible translations without inventing new ones, collaborative projects like Wikipedia and Wiktionary have been used, especially the Latin Wikipedia that has more than 100.000 entries (Vicipaedia, 2016) proved to be very useful.

Other typical challenges in creating a lexicon include homonymy and words that have no direct equivalent in the target language. They have been dealt with in the usual ways, either with multiple entries for homonymy or by paraphrase.

Our lexicon contains base forms and necessary grammatical information for all entries. Furthermore the Latin grammar contains morphological rules that automatically generate the full-form lexicon.

4. Morphology

This grammar contains a comprehensive implementation of Latin morphology. Latin is a language with strong inflection. Most of the lexical categories are inflected for several grammatical features and the inflection follows certain inflection classes, ie classes of words of the same category that construct word forms by a similar schema.

To get the paradigm, ie the set of all possible forms, for a lexicon entry, pattern matching is applied on the base form from the lexicon to find the appropriate inflection class which then is used to generate the missing forms. In the terminology of GF this is called a smart paradigm (Ranta, 2011, p. 82).

Thanks to the great regularity of the Latin morphology, it is possible to reduce the information needed for storing a lexicon entry to only one or just a few word forms. From these the whole paradigm of up to 260 forms[1] is inferred by the smart paradigm. Only for a few exceptions significantly more forms have to be listed.

The main idea for the implementation of morphology is to list all inflection forms using tables in GF that depend on grammatical features that are specific for the different word categories. An overview of the parametric features in this grammar can be seen in Table 1. The domains of these features are presented in Table 2.

First stem forms or other intermediate base forms are computed, to which the correct suffix according to the features is attached. These suffixes are mostly regular with only a few exceptions.

Nouns in Latin are inflected by case and number while they have an inherent gender. The Latin cases are given in Table 2. So nouns in Latin can have 12 different

[1] All possible verb forms include nominal and adjectival forms like gerund, gerundive and supine

Word class	Inherent	Parametric	No. of Inflection classes
Noun	Gender	Number, Case	5
Adjective		Degree, Gender, Number, Case	3
Verb (active)		Anteriority, Tense and Mood, Number, Person	4 regular, 4 deponent
Determiner	Number	Gender, Case	

Table 1: Inherent and parametric features for some lexical categories

Feature	Values
Gender	Feminine, Masculine, Neuter
Number	Singular, Plural
Case	Nominative, Genitive, Dative, Accusative, Ablative, Vocative
Degree	Positive, Comparative, Superlative
Anteriority	Anterior, Simultaneous
Tense and Mood	Present Indicative, Present Subjunctive, Imperfect Indicative, Imperfect Subjunctive, Future
Person	1, 2, 3

Table 2: Domains of the finite features

forms. These forms are created according to five declension classes.

The inflectional behaviour of adjectives is comparable to the one of nouns except that the inflected word forms depend on two further parameters, the gender and the comparison degree. Also there are only three declension classes for Latin adjectives but the comparison levels are sometimes formed using comparison adverbs instead of using morphology. These cases cannot be handled using simple inflection tables and have to be handled otherwise.

The most strongly inflected lexical class are verbs. The usual finite verb forms depend on several features (see Table 1).

Due to the requirements of the RGL to accommodate for the multilinguality, the usual tense system is split up in the tenses and anteriority forms as seen in the table. This tense system is based on the work of Reichenbach (1947, pp. 287-298) and can be uniquely mapped to the traditional Latin tenses. Additionally, there are forms derived from verbs that are used in a nominal or adjectival way like gerund or gerundive.

A special case are the so called deponent verbs, ie verbs that are used in active voice while their forms are similar to verb forms in passive voice. For these verbs the suffixes have to be adjusted accordingly. Also, since they use passive forms for active voice, they are missing the forms for passive voice (Bayer and Lindauer, 1994, p. 83).

Besides these three major word classes which have been presented here, we also implemented morphological rules for further word classes. These additional word classes consist mostly of different kinds of pronouns, as well as determiners.

5. Syntax

The function of the syntax rules is to assemble larger syntactic parts from smaller ones starting by lexical elements up to the sentence level.

Latin is well known for the flexibility in word order. So our challenge has been to figure out what parts should be already fixed in their position and what parts have to be kept separate to guarantee the necessary flexibility.

The GF construct we can use to keep parts apart as long as necessary are the records mentioned before. In a record we can keep multiple parts of a phrase separate until we fix their order.

Another important part of the syntactic rules is to guarantee that the agreement between the inherent features and the parametric features is kept sound. The GF tables, parameters and records makes this task quite straightforward.

The syntactic constructions that we have implemented start with different kinds of the usual noun and verb phrases that can be combined to more complex sentences. These include, besides basic declarative sentences, different forms of questions such as questions introduced by interrogative pronouns and Yes/No-questions. Also coordination on the adjective level is supported. Some of the missing parts include the numeral system, the handling of relative clauses. and coordination on other levels.

Finally, Latin is especially known for the free word order. But an analysis by Bamman and Crane (2006) shows, that there is for each period of use of the Latin language a strong tendency towards a specific order. Since a focus on the classic period of Latin language and literature seems most reasonable, we decided to use the predominant word order of this time, Subject-Object-Verb. Still the remaining five reasonable choices for word order can be used.

6. Implementation Status

The current state of the implementation covers a total of 530 out of 847 constructions defined in the abstract syntax of the RGL consisting of 475 lexical rules and 55 syntactic rules. The missing constructions consist of 92 lexical and 225 syntactic rules. The main work was done by one person over a period of about six month.

The aim of this work has been to explore the applicability of the GF grammar system to a both strongly inflectional and syntactic flexible language like Latin. We succeeded insofar that we were able to translate the morphological rules listed in a ordinary school grammar book into rules in Grammatical Framework as well as added rules that can

handle aspects of free word order. The outcome provides a base for both further research and applications as well as a general linguistic resource that is freely available.

7. Applications

Besides the pure focus on the implementation of this Latin grammar, we were also considering possible applications. For that two possible fields came to mind, firstly the application in Cultural Heritage and secondly to help people trying to learn Latin. It might contribute to give people ways to experience the rich history and culture in Europe that is quite often connected to the Latin language, may it be the Roman empire dominating the Classical antiquity or the Catholic church dominating the Medieval ages. In connection with other resources like the Epigraphic Database Heidelberg (Heidelberg, 2016), that collect Latin inscriptions and epigraphs, a translator for Latin inscriptions can be built.

On the other hand, the grammar might be used to exercise and learn Latin, eg for High-School or University students.

References

David Bamman and Gregory Crane. 2006. The design and use of a latin dependency treebank. In Jan Hajic and Joakim Nivre, editors, *Proceedings of the Fifth International Treebanks and Linguistic Theories Conference*, pages 67–78, Prag. Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University.

Karl Bayer and Josef Lindauer, editors. 1994. *Lateinische Grammatik*. C.C. Buchners Verlag, J. Lindauer Verlag, R. Oldenburg Verlag, 2. edition, auf der grundlage der lateinischen schulgrammatik von landgraf-leitschuh neu bearbeitete edition.

Epigraphic Database Heidelberg. 2016. Epigraphic database heidelberg. `http://edh-www.adw.uni-heidelberg.de/home/`. Online, accessed 7.5.2016.

Peter Ljunglöf. 2004. *Expressivity and Complexity of the Grammatical Framework*. Ph.D. thesis, Göteborgs University.

Aarne Ranta. 2009. The gf resource grammar library. *Linguistic Issues in Language Technology*, 2(2), December.

Aarne Ranta. 2011. *Grammatical Framework*. CSLL Studies in Computational Linguistics.

Hans Reichenbach. 1947. *Elements of Symbolic Logic*. The Macmillan Company.

Vicipaedia. 2016. Vicipaedia libera enceclopaedia. `https://la.wikipedia.org/wiki/Vicipaedia:Pagina_prima`. Online, accessed 4.4.2016.