PrinParLat: a resource of Latin principal parts

Matteo Pellegrini, Marco Passarotti, Francesco Mambrini, Giovanni Moretti
CIRCSE Research Centre, Università Cattolica del Sacro Cuore, Milano

1 Introduction

In this talk, we present PrinParLat, a free lexical resource documenting Latin verb inflection, making use of notions of theoretical morphology to provide rich information in a compact way.

Firstly, PrinParLat is a collection of principal parts. This notion was already used in traditional Latin dictionaries, where for each entry the citation form is accompanied by a set of forms from which the full paradigm can be inferred – e.g., the present active infinitive *amāre*, the first-person singular perfect active indicative *amāvī* and the perfect participle *amātum* for AMŌ 'love' in the Oxford Latin Dictionary – and it has recently been implemented in a principled fashion in theoretically grounded studies that investigate the implicative structure of paradigms with different approaches (Stump & Finkel, 2013; Bonami & Beniamine, 2016).

Secondly, two different layers of lexical units are used: each principal part is linked not only to the corresponding lexeme, but also to the corresponding flexeme(s). This distinction was introduced by Fradin & Kerleroux (2003) to account for cases of lexical items with different meanings but with the same form in all paradigm cells – e.g., for the French noun FILLE, there are two different lexemes (one for the meaning 'girl', one for the meaning 'daughter') that map to the same flexeme, as the wordforms are the same – and it has been recently applied (Bonami & Crysmann, 2018; Thornton, 2018) to the converse case of lexical items with the same meaning but different forms – i.e., to cases of overabundance; e.g., for the Italian noun ORECCHIO/A there are two different flexemes (one for the masculine forms *orecchio* SG and *orecchi* PL, one for the feminine forms *orecchia* SG and *orecchie* PL) that map to the same lexeme, as the meaning is the same ('ear').

Thirdly, regarding the inflectional behaviour of lexical items, information on the traditional conjugations of Latin verbs is provided. These can be considered as inflection "macro-classes" (Dressler, 2002; Beniamine et al., 2017), as they group items that are inflected similarly, but not identically – namely, they are inflected in the same way in imperfective wordforms, but not in the other ones. Furthermore, lexical items are also classifed according to their fine-grained inflection "micro-class", grouping together the ones that are inflected in the same way across the whole paradigm. These micro-classes are identified in an abstractive fashion (Blevins, 2016), by inspecting the alternation patterns that occur between all possible combinations of the listed principal parts.

2 The resource

The data of the resource are taken from the database of a morphological analyzer of Latin, Lemlat (Passarotti et al., 2017). The stems and endings reported there for verbs have been used to generate the full wordforms that we choose as principal parts. To restrict the remarkable time span covered by the Latin language, we only select about 8,000 entries that come from dictionaries of Classical Latin, thus excluding Medieval Latin verbs recorded in the database.

The resource is structured as a relational database, using the tables and columns defined in an emerging standard format for paradigmatic lexicons, Paralex. The core part is the forms table (1a), where for each principal part, we provide information on the cell it fills, the lexeme it belongs to, and its form. Due to the unclear epistemological status of the actual pronunciation of Classical Latin, that can only be reconstructed, we provide orthographic transcriptions, rather than phonetic/phonological ones. We follow the traditional usage of Latin grammars and dictionaries in selecting PRF.ACT.IND.1.SG and PRF.PTCP.NOM.N.SG as principal parts from which perfective wordforms and forms displaying Aronoff (1994)'s Third Stem can be inferred, respectively (e.g., *amāvī*; *amātum*, for the verb meaning 'love'). We depart from the tradition in selecting PRS.ACT.INF and FUT.ACT.IND.3.SG – rather than the citation form PRS.ACT.IND.1.SG – as the principal parts from which imperfective wordforms can be inferred. This is due to the fact that the first-person singular is actually poorly informative on the content of other cells, as it neutralizes the opposition between 1st and 3rd conjugation verbs. Furthermore, an additional principal part is provided, namely FUT.PTCP.NOM.N.SG, to be able to infer future participle forms also in the few cases in which they display a stem different than the one of perfect participle forms. Additional principal parts are also needed for defective lexemes: for instance, we use the corresponding passive forms for deponent verbs that lack the active ones.

Additional information is provided in separate tables. For instance, regarding cells, we rely on the traditional description of the Latin verbal system, as documented in the features-values table (1c). However, in the cells table (1b), the corresponding notation in the UniMorph format is given (McCarthy et al., 2020), thus allowing for a mapping to a more theoretically grounded and interlinguistically consistent vocabulary.

Furthermore, we introduce custom tables and columns, not defined in the Paralex standard, but required by the characteristics of our data. In the forms table, an additional column for flexemes is needed, to allow for the expression of both the layers of lexical units described in Section 1. Consequently, an additional table (1d) is also introduced to provide information on flexemes. Inflection classes are assigned to flexemes (rather than lexemes), as lexical items identified according to their form (rather than their meaning) appear to be the appropriate locus to encode a classification based on form. Each flexeme is associated both to its traditional conjugation, expressed with the labels used in the LiLa Knowledge Base of interoperable resources for Latin (Passarotti et al., 2020) – on which see below, Section 3 – and to an index corresponding to its fine-grained inflection micro-class. Micro-classes are automatically inferred from data, using the Qumin toolkit (Beniamine, 2018) to extract alternation patterns between all the possible combinations of principal parts for each flexeme, and grouping together flexemes that share the same set of patterns, as documented in the tables in (1e-f).

3 Conversion to RDF and linking to the LiLa Knowledge Base

Having PrinParLat released in the Paralex standard format will make it interoperable with other Paralex lexicons. However, for Latin a wealth of other resources of different kinds is also available, and some of them provide pieces of information that can integrate the ones explicitly recorded in our resource. For instance, in large textual resources like the LASLA corpus (Denooz, 2004) we can find information on the frequencies of wordforms, which is particularly useful as they are generated regardless of their actual attestation in our resource. Lexical resources focusing on other topics can prove useful as well: e.g., a derivation lexicon like Word Formation Latin (Litta & Passarotti, 2020) can give us information on which of the items of our resource are linked by a word formation relation, and how this influences their inflectional behaviour.

To guarantee interoperability with such resources, a richer integration is needed, that can be achieved by means of Semantic Web technologies and standards. Indeed, many of the resources available for Latin have already been made interoperable by connecting them to the LiLa Knowledge Base (cf. Section 2), that follows the RDF data model, where knowledge is

(a) The forms table

form_id	lexeme	cell	orth_form	flexeme
192	a0105	prs.act.inf	ablauare	a0105
193	a0105	prs.act.inf	ablauere	a0105_2
190	a0105	fut.act.ind.3.sg	ablauabit	a0105
191	a0105	fut.act.ind.3.sg	ablauet	a0105_2

(b) The cells table

cell_id	unimorph
prs.act.inf	V;NFIN;ACT;IPFV

(c) The features-values table

value_id	value_label	feature
prs	present	tense-aspect
act	active	voice
inf	infinitive	verbform

(d) The flexemes table

flexeme_id	inflection_class	lila:Lemma	lila:hasInflectionType
a0105	20	86938	v3r
a0105_2	21	86939	v1r

(e) Mapping inflection classes-patterns

id	inflection_class	pattern
113	20	1
116	21	0

(f) The patterns table

pattern_id	pattern_alternation	cell_left	cell_right
0	re ⇌ bit	prs.act.inf	fut.act.ind.3.sg
1	er_ ⇌ _t	prs.act.inf	fut.act.ind.3.sg

Table 1: The data of PrinParLat

represented in terms of triples that connect a "subject" to an "object" through a "property", items ("individuals") are assigned to "classes", and sub-class and sub-property relations are established between them to describe their characteristics. Following the principles of the Linguistic Linked Open Data paradigm (Cimiano et al., 2020), already existing vocabularies are reused whenever possible – e.g., the OntoLex-Lemon model for lexical resources (McCrae et al., 2017). New classes and properties are introduced whenever necessary. Among the extensions of the LiLa ontology, the crucial one is the class `lila:Lemma`, defined as a subclass of `ontolex:Form`: the core of the Knowledge Base is the Lemma Bank, a large collection of citation forms of Latin words; interoperability is achieved by linking tokens of textual resources and entries of lexical resources to their citation form.

To make our resource interoperable with those already included in the Knowledge Base, we need to also release it in RDF. To do that, Paralex also provides an ontology where tables and columns defined in the standard are mapped to RDF classes and properties, respectively. However, we also need i) to extend this vocabulary to be able to model tables and columns of our resource that are not defined in the standard, and ii) to specify how the conversion should be implemented. Linking to the Knowledge Base can then be performed by connecting each flexeme to its `lila:Lemma` in the Lemma Bank, as shown in the flexemes table in (1d).

4 Conclusions and future work

PrinParLat lists the principal parts of Latin verbal (f)lexemes and provides fine-grained information on their inflectional behaviour. Putting all these pieces of information together, it is straightforward to obtain a full lexicon listing all the inflected wordforms of Latin verbs, by performing simple string replacements compatible with the relevant inflection (micro-)class in each of the other cells. The instructions to obtain it can be coded in RDF too, using the vocabulary of the emerging module for the treatment of morphological information in OntoLex lexicons (Chiarcos et al., 2022). Wide-scope interoperabilty of such a lexicon would be guaranteed with i) other paradigmatic lexicons, thanks to the adoption of the Paralex standard format; ii) other lexical resources that use the OntoLex vocabulary, thanks to the explicit mapping between the Paralex standard and OntoLex provided in the Paralex ontology; iii) resources of other kind (e.g. corpora), thanks to its release as RDF data linked to the LiLa Knowledge Base.

References

Aronoff, Mark. 1994. *Morphology by itself: Stems and inflectional classes*, vol. 22. MIT press.

Beniamine, Sacha. 2018. *Classifications flexionnelles. Étude quantitative des structures de paradigmes*: Université Sorbonne Paris Cité-Université Paris Diderot (Paris 7) dissertation.

Beniamine, Sacha, Olivier Bonami & Benoît Sagot. 2017. Inferring inflection classes with description length. *Journal of Language Modelling* 5(3). 465–525.

Blevins, James P. 2016. *Word and paradigm morphology*. Oxford University Press.

Bonami, Olivier & Sacha Beniamine. 2016. Joint predictiveness in inflectional paradigms. *Word Structure* 9(2). 156–182.

Bonami, Olivier & Berthold Crysmann. 2018. Lexeme and flexeme in a formal theory of grammar. In *The lexeme in descriptive and theoretical morphology*, 175–202. Language Science Press.

Chiarcos, Christian, Katerina Gkirtzou, Fahad Khan, Penny Labropoulou, Marco Passarotti & Matteo Pellegrini. 2022. Computational Morphology with OntoLex-Morph. In *Proceedings of the 8th workshop on linked data in linguistics within the 13th language resources and evaluation conference*, 78–86.

Cimiano, Philipp, Christian Chiarcos, John P McCrae & Jorge Gracia. 2020. *Linguistic Linked Data*. Springer.

Denooz, Joseph. 2004. Opera Latina: une base de données sur internet. *Euphrosyne* 32. 79–88.

Dressler, Wolfgang U. 2002. Latin inflection classes. In *Theory and description in Latin linguistics*, 91–110. Brill.

Fradin, Bernard & Françoise Kerleroux. 2003. Troubles with lexemes. In *Topics in Morphology. Selected papers from the Third Mediterranean Morphology Meeting*, 177–196. IULA-Universitat Pompeu Fabra (Barcelona).

Litta, Eleonora & Marco Passarotti. 2020. (When) inflection needs derivation: a word formation lexicon for Latin. In *Lemmata Linguistica Latina. Words and Sounds*, 224–239. De Gruyter.

McCarthy, Arya D, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J Mielke, Garrett Nicolai, Miikka Silfverberg et al. 2020. UniMorph 3.0: Universal Morphology. In *Proceedings of The 12th language resources and evaluation conference*, 3922–3931. European Language Resources Association.

McCrae, John P, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar & Philipp Cimiano. 2017. The Ontolex-Lemon model: development and applications. In *Proceedings of eLex 2017 conference*, 19–21.

Passarotti, Marco, Marco Budassi, Eleonora Litta & Paolo Ruffolo. 2017. The Lemlat 3.0 package for morphological analysis of Latin. In *Proceedings of the NoDaLiDa 2017 workshop on processing historical language*, 24–31.

Passarotti, Marco, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo & Rachele Sprugnoli. 2020. Interlinking through lemmas. The lexical collection of the LiLa Knowledge Base of linguistic resources for Latin. *Studi e Saggi Linguistici* 58(1). 177–212.

Stump, Gregory & Raphael A Finkel. 2013. *Morphological typology: From word to paradigm*, vol. 138. Cambridge University Press.

Thornton, Anna M. 2018. Troubles with flexemes. In *The lexeme in descriptive and theoretical morphology*, 303–321. Language Science Press.