STUDI E SAGGI LINGUISTICI

LVIII (1) 2020

rivista fondata da

Tristano Bolelli

Edizioni ETS

SSL

STUDI E SAGGI LINGUISTICI

www.studiesaggilinguistici.it

SSL electronic version is now available with OJS (Open Journal Systems)
Web access and archive access are granted to all registered subscribers

Abbonamento, compresa spedizione
individuale, Italia € 50,00
individuale, Estero € 70,00
istituzionale, Italia € 60,00
istituzionale, Estero € 80,00
Bonifico su c/c Edizioni ETS srl
IBAN IT 21 U 03069 14010 100000001781
BIC BCITITMM
Causale: Abbonamento SSL

Subscription, incl. shipping
individual, Italy € 50,00
individual, Abroad € 70,00
institutional, Italy € 60,00
institutional, Abroad € 80,00
Bank transfer to Edizioni ETS srl
IBAN IT 21 U 03069 14010 100000001781
BIC BCITITMM
Reason: Subscription SSL

L'editore non garantisce la pubblicazione prima di sei mesi dalla consegna in forma definitiva di ogni contributo.

Registrazione Tribunale di Pisa 12/2007 in data 20 Marzo 2007
Periodicità semestrale
Direttore responsabile: Alessandra Borghini

ISBN 978-884675901-6
ISSN 0085 6827

Sommario

Introduction

Marco Passarotti

1. *Preliminary remarks*

Lemmatization is a fundamental task in the linguistic annotation of both lexical and textual resources, lemmas serving as gateways to lexical entries in dictionaries, glossaries and lexica, as well as to single occurrences of lexical items in textual corpora.

Since the early days of linguistic computing, as corpora grew in size so did the need for not only concordances, but 'lemmatized' concordances, to automatically investigate textual data. In 1949, Father Roberto Busa's pioneering machine-readable corpus, the Index Thomisticus (Busa, 1974-1980), was specifically conceived to provide scholars with a lemmatized concordance of the *opera omnia* of Thomas Aquinas.

Regrettably, however, the publication of computerized concordances with lemmatization has not been common practice[1]. Such practice was mainly due to the labor-intensive nature of the work of lemmatization, which relies on contextual analysis to disambiguate word forms to which more than one lemma and/or part of speech (PoS) can be assigned. However, the availability of large annotated corpora for many languages and the explosion of the empirical paradigm in natural language processing (NLP) in the nineties made it possible to develop stochastic lemmatizers and PoS taggers able to provide high accuracy rates[2]. An overview of the current state of the

[1] In an article published in 1983, Father Busa explicitly complained about the widespread habit of producing unlemmatized concordances: «mi lamento che non si fa se non produrre concordanze troppo spesso ahimé nemmeno lemmatizzate, che poi nessuno studia» (Busa, 1983: § 7.4). English translation by Philip Barras (Nyhan and Passarotti, 2019: 142): "I am sorry that all that happens is the production of concordances, which, alas, too often are not even lemmatized, and which then nobody studies".

[2] There are two main paradigms in NLP, namely the rule-based (or intuition-based) paradigm and the empirical (or data-driven) paradigm. Rule-based tools are built around a set of (manually-

art in the field can be found in the results of the recent *CoNLL* 2018 Shared Task (Zeman *et al.*, 2018). Although the shared task was focused on learning and evaluating dependency parsers for a large number of languages based on test sets adhering to the unified Universal Dependencies (*UD*) annotation scheme[3], results on lemmatization and PoS tagging were also provided. The ranking of participating tools shows that the best system for lemmatization achieves a macro-averaged score of 91.24 of correctly assigned lemmas over 82 test treebanks in 57 languages, while the winner system for PoS tagging reaches a score of 90.91 (Zeman *et al.*, 2018: 10).

Thanks to the availability of huge amounts of (raw) linguistic data, and of computers powerful enough to process them, several machine learning techniques can now achieve good accuracy rates in various NLP tasks with both supervised and unsupervised methods for many languages. Nevertheless, linguistic annotation is still necessary for those (historical) languages that cannot rely on billion-word text collections. Lemmatization, in particular, is the first level of lexical categorization in annotation; by collecting all the textual occurrences of a lexical item under the same citation form, it provides essential support to information retrieval. And yet, the patchy lemmatization evaluation of most of the Latin text collections currently available severely impacts information retrieval. Indeed, even if enhanced with regular expressions, string- or character-matching queries on an unlemmatized corpus, risk generating both low precision (many false positives) *and* low recall (many false negatives). Moreover, owing to the philological tradition in Classics and the limited availability of texts in Latin, community expectations of the quality of both raw data and annotations is very high. For most languages, and particularly Latin, such quality is hardly achievable through automation alone.

The high degree of diversity of Latin texts introduced by the language's wide diachrony and diatopy, makes it difficult to build one-size-fits-all NLP tools able to sustain high performance on texts of different genres, eras and

crafted) linguistic rules and tend to be language-dependent. In contrast, data-driven tools, use (language-independent) machine learning techniques (based on different kinds of statistical methods) to create NLP models that are trained on a set of data provided by linguistic resources, such as (annotated) corpora. While the rule-based paradigm was predominant in the NLP community until the nineties, the empirical paradigm has since taken over thanks to the increasing availability of linguistic data in digital format.

[3] Universal Dependencies is a community-driven initiative, which aims to build a collection of syntactically annotated corpora (called 'treebanks') for several languages following a common dependency-based annotation style (*https://universaldependencies.org*).

origin, particularly when these belong to a domain other than that of the training data. In this respect, the results of the recent evaluation campaign of NLP tools for Latin *EvaLatin* (Sprugnoli *et al.*, 2020) show a decrease of an average 5-10 points on the lemmatization accuracy of cross-genre and cross-time data. The winning system, trained on Classical Latin data, reaches an accuracy rate of 96.19 on Classical Latin but drops to 91.01 on cross-time data and to 87.13 on cross-genre data (Sprugnoli *et al.*, 2020: 107).

Another issue affecting Latin lemmatized text collections (those counting a few million words) is their use of different criteria, tag sets and formats to assign and record both lemmas and PoS. This heterogeneity prevents corpora from interacting with one another without time-consuming and potentially lossy conversion processes, and from being used to build a single, common training set for the development of stochastic NLP models. The four Latin treebanks available in the *UD* format are no exception[4]. While employing the same syntactic annotation style and the same tag set for PoS and morphological features, their lemmatization and PoS tagging criteria diverge in a number of aspects, for instance the treatment of participles.

Given that Latin is a dead language and that textual production today is limited to a few texts only (notably, by the Vatican State), the lemmatization of the entire corpus of Latin texts available seems, at least in principle, possible. Such an objective is, however, difficult to achieve in the short term, not only because of the current limitations in NLP for Latin, but also because of the amount (and, thus, diversity) of the data to process. Indeed, the size of the entire Latin corpus might not qualify as Big Data, yet it is considerable, mostly as a consequence of Latin's *lingua franca* role played all over Europe up until the 1800s (Leonhardt, 2009). The Open Greek and Latin project[5], estimated Ancient Greek and Latin production surviving from Antiquity through 600 AD at approximately 150 million words, and from an analysis of 10,000 books written in Latin available from *archive.org*, the project also identified over 200 million words of post-Classical Latin. This body of text does not include the sizeable Neo-Latin literature, that is, texts dating

[4] The four Latin treebanks available in *UD* are the Index Thomisticus Treebank (Cecchini *et al.*, 2018), which collects a selection of the works of Thomas Aquinas; the Latin Dependency Treebank (Bamman and Crane, 2006) of texts belonging to the Classical era; the *PROIEL* corpus (Haug and Jøhndal, 2008), featuring the oldest extant versions of the New Testament in Indo-European languages and a number of Latin texts from both the Classical and the Late era, and the Late Latin Charter Treebanks (Korkiakangas and Passarotti, 2011), based on charters of the 8th-9th century AD.

[5] Cf. *https://www.dh.uni-leipzig.de/wo/projects/open-greek-and-latin-project/*.

from the age of Petrarch (1304-1374) to the present day[6]. The predominance of Latin in early modern Europe is evidenced by the Universal Short Title Catalogue[7]: out of almost 750,000 bibliographical entities (dating between the invention of print and 1650) catalogued in 8,500 memory institutions, more than 280,000 are in Latin, followed, in second place, by French with approximately 100,000 entries.

2. *Aims and contents of this Special Issue*

Recognizing the relevance of lemmatization for Latin linguistic resources, this special issue of *Studi e Saggi Linguistici* is devoted to 'Current Approaches in Latin Lemmatization'.

In collecting a selection of articles about the strategies and methods in lemmatization and PoS tagging adopted in a number of linguistic resources and NLP tools for Latin, this special issue aims to assess the state of the art in this area with a view to understanding the problems raised by resource interoperability. Indeed, domain experts are faced with an increasing need to harmonize (meta)data differences for the benefit of the wider Humanities community.

The special issue is divided into three sections. The first two sections feature three papers each, and deal, respectively, with issues of lemmatization and with lemmatization tools. These inform the third section, which includes a paper specifically on the pursuit of interoperability through lemmatization.

2.1. *Issues of lemmatization in Latin corpora*

The first section of the special issue addresses lemmatized Latin corpora comprising texts of different eras, origin and type. Celano's article, for instance, discusses issues of lemmatization of Classical literary Latin in a dependency treebank; Marotta *et al.* introduce a corpus of non-literary Latin inscriptions, letters and tablets from various Roman provinces written between the 4th century BC and the 6th century AD. Finally, Korkiakangas

[6] The most comprehensive collection of Neo-Latin texts, the CAMENA corpus (*http://mateo.uni-mannheim.de/camenahtdocs/camena_e.html*), counts about 50 million words.

[7] Cf. *https://www.ustc.ac.uk/about*.

discusses questions of lemmatization in a syntactically annotated corpus of original 8th-9th century AD charters from Central Italy.

The article by Giuseppe Celano (*Lemmatization and morphological analysis for the Latin Dependency Treebank*) highlights one of the main issues related to lemmatization, namely the harmonization of the different annotation criteria and tag sets used by resources and tools today. The paper provides an overview of the challenges raised by Latin lemmatization and PoS tagging, focusing on the workflow of morphological annotation adopted for the Latin Dependency Treebank[8]. The author discusses the issues concerning the choice of the lemma as the canonical form representing the inflectional paradigm of a word, and the question of the set of the PoS, more specifically the treatment of participles, nominalized adjectives and gerundives/gerunds. These problems are presented in light of a wider discussion on the differences between the Latin lemmatizers and morphological analyzers available.

The paper by Marotta *et al.* (*CLaSSES: Orthographic variation in non-literary Latin*) introduces *CLaSSES* (Corpus for Latin Sociolinguistic Studies on Epigraphic textS), an annotated corpus of approximately 3,500 non-literary Latin texts (epigraphs, writing tablets, letters)[9]. The texts cover a wide diachronic span (6th century BC-7th century AD) and show a diverse distribution of their places of provenance, including four provinces of the Roman Empire, namely Rome (and Italy), Roman Britain, Egypt and the Eastern Mediterranean, and Sardinia. The non-literary nature of the *CLaSSES* texts provides substantial empirical evidence of Latin's orthographic variation through time and space. The wide range of annotations, described here in great detail, prove particularly useful in this regard and support both qualitative and quantitative orthographic investigations. Indeed, besides the standard layers of linguistic and extra-linguistic annotation (such as lemmatization and textual typology), the corpus also carefully annotates misspellings with the objective of collecting and classifying non-classical variant forms according to the variation phenomenon shown. In adopting a strictly descriptive approach to the annotation of (ortho-) graphic phenomena, each spelling variant is labelled as 'non-classical' and is associated to its corresponding classical standard form. Another distinctive feature of *CLaSSES* is that a graphic form category is assigned to each word form, like, for instance, abbreviation, incomplete word and lacuna.

[8] Cf. *https://perseusdl.github.io/treebank_data/*.
[9] Cf. *http://classes-latin-linguistics.fileli.unipi.it*.

Providing this kind of annotation proves to be particularly helpful, as the texts collected in *CLaSSES* are originally written on supports whose conservation status often results in faint or missing letters.

Timo Korkiakangas (*Theoretical and pragmatic considerations on the lemmatization of non-standard Early Medieval Latin charters*) tackles the important question of lemmatization of non-standard late Latin. The article discusses the theoretical and practical questions related to the lemmatization of the Late Latin Charter Treebanks (*LLCT*), a set of three dependency treebanks of Early Medieval Latin documentary texts (charters) written in Italy between 714 and 1000 AD. The paper focuses on the two guiding principles of the lemmatization of the *LLCT*: the evolutionary principle and the parsimony principle. The evolutionary principle aims at reducing linguistic variants brought about by language evolution to their standard-Latin 'ancestor' forms. The article details the different types and origin of variants found in the *LLCT*, discussing the treatment of variation in inflectional endings, proper names, loans from other languages (mostly, Germanic), Late Latin neologisms, non-derived Early Medieval formations of uncertain origin and mistaken words. The parsimony principle states that lemmas do not have to be unnecessarily multiplied. The paper focuses on the lemmatization of forms that have changed inflectional properties, claiming that they must be analyzed under the same lemmas rather than creating new, separate lemmas. Such a solution fits the properties of later written Latin, where «borders between declensions, conjugations, and genders had become increasingly permeable in several morphophonological contexts [...], without implying a change in meaning» (p. 86).

2.2 *Automatic lemmatization of Latin*

The second section of the special issue includes papers about automatic lemmatization of Latin, presenting NLP tools that make use of different techniques and approaches. While the lemmatizer introduced by Verkerk *et al.* is based on a large collection of textual data, which makes it possible to achieve high accuracy rates despite the simple statistical model adopted by the tool, the article by Mehler *et al.* focuses on the role played by lexical data in automatic lemmatization. Finally, on the opposite to the approach of Verkerk *et al.,* is that described by Burns, who introduces a method that makes use of a series of sub-lemmatizers to overcome the limited amount of empirical evidence supporting automatic lemmatization for Latin.

The paper by Verkerk *et al.* (*L.A.S.L.A. and Collatinus: A convergence in lexica*) presents the lemmatization provided by the large Opera Latina corpus developed since the sixties at the *L.A.S.L.A.* laboratory in Liège (Laboratoire d'Analyse Statistique des Langues Anciennes) and describes the Collatinus lemmatizer, which is strictly related to *Opera Latina*[10]. The authors detail the structure of the files of the corpus, the tokenization procedure, the lemmatization criteria, as well as the layer of morphological annotation and PoS tagging. The paper describes the functionalities of the *L.A.S.L.A.* Encoding Initiative interface, which allows users to check the results of an out-of-context procedure of automatic tokenization, lemmatization and morphological analysis. The two interfaces available to query the (meta)data of Opera Latina are also presented. As for Collatinus, the paper provides an overview of the linguistic analysis performed by the tool, which, besides lemmatization and morphological analysis, also assigns lemmas their definition(s) – taken from four dictionaries of Latin[11] –, as well as their metrical structure. The authors detail the process of segmentation of the input forms and discuss a number of issues concerning the treatment of the enclitics, assimilations, contractions and graphical variants. A section of the paper deals with the lexical basis of Collatinus (counting some 77,000 lemmas) and its extension to lemmatize a large Medieval corpus. Collatinus also performs automatic disambiguation of ambiguous lemmatizations through a Hidden Markov Model statistical tagger, trained on the Opera Latina corpus. The paper concludes with a comparison between the lemmatization process pursued to prepare the *L.A.S.L.A.* files, which requires that a scholar select the correct analysis from a set of possibilities, and that of the statistical tagger, where the role of the human annotator is to check the analysis proposed by the tool.

Mehler *et al.* (*The Frankfurt Latin Lexicon. From morphological expansion and word embeddings to SemioGraphs*) present the Frankfurt Latin Lexicon (*FLL*)[12]. The *FLL* is a morphological lexicon for Medieval Latin covering the period between 400 and 1500 AD and supporting both the automatic lemmatization of Latin texts (with the Text-technology Lab Latin Tagger) and the post-editing of the lemmatization process. The paper details the features of the *FLL*, focusing on its layers of lexical annotation,

[10] Cf. *http://web.philo.ulg.ac.be/lasla/*.

[11] Georges and Georges (1913-1918), Gaffiot (1934), Lewis and Short (1966), and the *Dictionnaire Latin-Français* by Gérard Jeanneau, Jean-Paul Woitrain and Jean-Claude Hassid available at *https://www.prima-elementa.fr/Dico.htm*.

[12] Cf. *https://www.comphistsem.org/lexicon0.html*.

the treatment of multi-word units and a tool to create all of the inflected forms for newly entered lemmas. A section of the paper is dedicated to the comparison of a number of lemmatizers trained on different Latin corpora and evaluated against both the *PROIEL* corpus and the Capitularies corpus, the latter produced by the Text-technology Lab in Frankfurt as a reference for Medieval Latin processing. As well as describing an extension of the *FLL* obtained through word embeddings, the paper stresses the need to use these in a stratified manner dependent on contextual parameters, such as genre and authorship, so as to represent the different (or similar) use of a word according to the parameters chosen. The authors present a series of graphical visualizations of their results, which are in turn used to perform a historical semantics analysis of three Latin words (*conclusio* "conclusion", *excommunico* "to communicate" and *pater* "father"). By comparing the results of a computational approach with those of traditional scholarship, these three case studies demonstrate the promise and need for an interaction between the 'two cultures' (Snow, 1959). In addition, the need to build word embeddings on smaller sets of data selected by genre and author rather than on large and generic collections of texts reflects a general issue related to the computational processing of Latin texts, i.e. the high degree of variation in the data used to train NLP tools or to feed visualizations to support claims grounded in distant reading techniques.

In his paper, entitled *Ensemble lemmatization with the Classical Language Toolkit*, Patrick Burns touches upon the issue of the narrow set of linguistic resources available for historical languages in support of lemmatization. The paper presents a solution called 'ensemble lemmatization', which consists of a series of sub-lemmatizers to limit the output to a single probable lemma or group of probable lemmas. The ensemble lemmatizer is developed for the Classical Language Toolkit, a widely used Python framework supporting NLP for historical languages[13]. The author shows the flexibility and extensibility of ensemble lemmatization. The user, in fact, is given a great degree of customization over the construction process of the lemmatizer, and the lemmatizer itself can use a wide range of data sources, including lexica, sentence-level training data, lists of regular expression patterns, as well as the output of other lemmatizers. Flexibility and extensibility are strictly related to modularity, licensing the author to describe ensemble lemmatization as philological. According to Burns, the multiple-pass tagging strategy based on dif-

[13] Cf. *http://cltk.org*.

ferent resources pursued by his lemmatizer reflects «established disciplinary practices for disambiguating words», namely «the decoding strategies of the philologically trained reader of historical texts» (p. 168). Such reference to traditional practices of (manual) lemmatization may sound strange in times ruled by deep learning techniques, where the size of the unsupervised training data matters more than steady annotation and strong linguistic expertise. And yet, the strict connection between century-long practices and new tools for automatic NLP is just what is peculiar of the application of such tools to historical languages, which lack both native speakers and, most importantly, large amounts of linguistic data. Once again, such a connection insists on the exchange and collaboration between historical and computational linguists.

2.3. *Interlinking linguistic resources for Latin through lemmatization*

As previously mentioned, today, many valuable linguistic resources for Latin remain unused (if not unknown), partially owing to the different lemmatization criteria they adopt. While common to many languages[14], the issue of resource interoperability in Latin lies at the heart of the *LiLa* project[15], introduced here by Passarotti *et al*.

Their article, entitled *Interlinking through lemmas. The lexical collection of the LiLa Knowledge Base of linguistic resources for Latin*, details the architecture supporting *LiLa*'s goal to overcome the lack of interoperability between Latin resources with the creation of a Knowledge Base based on the Linked Data paradigm, i.e. a collection of interlinked data sets described with the same vocabulary of knowledge description. Seeing as textual and lexical resources in the Knowledge Base interact through lemmatization, the core of *LiLa* consists of a large collection of Latin lemmas: interoperability is achieved by linking all those entries in lexical resources and tokens in corpora that point to the same lemma. The *LiLa* Knowledge Base does not force one single lemmatization style on the different corpora and tools it includes but harmonizes these into a dynamic Linked Data ecosystem. Like other papers in this volume, this article too discusses the problem posed by

[14] See the Linguistic Linked Open Data cloud of interoperable linguistic resources (*https://linguistic-lod.org*).

[15] Cf. *https://lila-erc.eu*. The project *LiLa: Linking Latin. Building a Knowledge Base of Linguistic Resources for Latin* has received funding from the European Research Council (*ERC*) under the European Union's Horizon 2020 research and innovation programme – Grant Agreement No. 769994.

different lemmatization strategies, focusing on the solutions found in *LiLa* to reconcile differences, particularly with regard to the various forms of the lemma and lemmatization criteria. *LiLa*'s underlying ontology, built as an extension of a number of existing (and *de facto* standard) ontologies, serves to represent the lemma bank and to ensure that resources in *LiLa* are compatible with other Linked (Open) Data resources. The paper illustrates how a lemma and its connected information are stored in the *LiLa* data structure and the inclusion in the Knowledge Base of a *UD*-compliant dependency treebank by way of example.

3. *Conclusion*

Seventy years of linguistic computing and steady work on the development of machine-readable linguistic resources (not to mention centuries of manual work on paper) notwithstanding, no general consensus has yet been reached on common lemmatization criteria, methods, formats and tag sets for Latin, let alone other languages, be those modern, ancient or historical. Such a predicament cannot be easily overcome by imposing one further, 'standard' set of best practices and rules for lemmatization; any such attempt would fail for the simple reason that lemmatization is not a black-or-white issue. After all, the different approaches adopted by corpora, dictionaries, glossaries and lexica are typically well motivated and supported by the individual projects' theoretical traditions and objectives.

By providing an overview of the various lemmatization processes and criteria applied in a number of linguistic resources and NLP tools for Latin, this special issue seeks to highlight their differences and commonalities, and points to interoperability as the necessary, nay, urgent, next step. Indeed, an efficient interaction of lemmatized linguistic resources can only be achieved in a dynamic ecosystem as that made possible by the Linked Data framework.

Acknowledgements

I am greatly thankful to Giovanna Marotta for suggesting the idea of this special issue to me, and for her continuous support and advice. Many thanks also to Francesco Rovai for the details of the review process and to Greta Franzini for her helpful suggestions.

References

Bamman, D. and Crane, G. (2006), *The design and use of a Latin dependency treebank*, in Hajič, J. and Nivre, J. (2006, eds.), *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories*, Institute of Formal and Applied Linguistics, Prague, pp. 67-78.

Busa, R. (1974-1980), *Index Thomisticus: sancti Thomae Aquinatis operum omnium indices et concordantiae, in quibus verborum omnium et singulorum formae et lemmata cum suis frequentiis et contextibus variis modis referuntur quaeque consociata plurium opera atque electronico IBM automato usus digessit Robertus Busa SJ.*, Frommann / Holzboog, Stuttgart / Bad Cannstatt.

Busa, R. (1983), *Trent'anni d'informatica su testi: a che punto siamo? Quali spazi aperti alla ricerca?*, in CILEA (1983, a cura di), *Atti del Convegno su 'L'Università e l'evoluzione delle Tecnologie Informatiche' (Milano 14-16 Marzo 1983)*. Vol. 1, CILEA, Milano, §§ 7.1-7.4.

Cecchini, F.M., Passarotti, M., Marongiu, P. and Zeman, D. (2018), *Challenges in converting the Index Thomisticus Treebank into Universal Dependencies*, in De Marneffe, M.C., Lynn, T. and Schuster, S. (2018, eds.), *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, Association for Computational Linguistics, Bruxelles, pp. 27-36.

Gaffiot, F. (1934), *Dictionnaire illustré Latin-Français*, Librairie Hachette, Paris.

Georges, K.E. and Georges, H. (1913-1918), *Ausführliches Lateinisch-Deutsches Handwörterbuch*, Hahn, Hannover.

Haug, D.T.T. and Jøhndal, M. (2008), *Creating a parallel treebank of the old Indo-European Bible translations*, in Sporleder, C. and Ribarov, K. (2008, eds.), *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, European Language Resources Association (ELRA), Paris, pp. 27-34.

Korkiakangas, T. and Passarotti, M. (2011), *Challenges in annotating medieval Latin charters*, in «Journal for Language Technology and Computational Linguistics», 26, 2, pp. 103-114.

Leonhardt, J. (2009), *Latein. Geschichte einer Weltsprache*, C.H. Beck, München.

Lewis, C.T. and Short, C. (1966), *A Latin Dictionary. Founded on Andrews' edition of Freund's Latin dictionary*, Clarendon Press, Oxford.

Nyhan, J. and Passarotti, M. (2019, eds.), *One Origin of Digital Humanities. Fr. Roberto Busa in His Own Words*, Springer International Publishing, Cham.

Snow, C.P. (1959), *The Rede lecture, 1959*, in Snow, C.P. (1959, ed.), *The Two Cultures: and a Second Look*, Cambridge University Press, Cambridge, pp. 1-22.

Sprugnoli, R., Passarotti, M., Cecchini, F.M. and Pellegrini, M. (2020), *Overview of the EvaLatin 2020 evaluation campaign*, in Sprugnoli, R. and Passarotti, M. (2020, eds.), *Proceedings of the LT4HALA 2020 Workshop - 1st Workshop on Language Technologies for Historical and Ancient Languages, satellite event to the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, European Language Resources Association (ELRA), Paris, pp. 105-110.

Zeman, D., Hajič, J., Popel, M., Potthast, M., Straka, M., Ginter, F., Nivre, J. and Petrov, S. (2018), *CoNLL 2018 Shared task: Multilingual parsing from raw text to Universal Dependencies*, in Zeman, D., Hajič, J., Popel, M., Straka, M., Nivre, J., Ginter, F. and Petrov, S. (2018, eds.), *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Association for Computational Linguistics, Bruxelles, pp. 1-21.

Marco Passarotti
Facoltà di Scienze Linguistiche e Letterature Straniere
Università Cattolica del Sacro Cuore
Largo Gemelli 1
20123 Milano (Italy)
marco.passarotti@unicatt.it

Saggi

SSL

Lemmatization and morphological analysis for the Latin Dependency Treebank

Giuseppe G.A. Celano

Abstract

The present article presents some challenges posed by lemmatization and PoS tagging of Latin, with reference to the ongoing work to revise the Latin Dependency Treebank. Current options available for lemmatization and morphological analysis of Latin are reviewed and discussed. The pipeline to annotate the morphological layer of the Latin Dependency Treebank is shown to consist in three main steps: (i) tokenization/sentence split, which is performed via a documented rule-based algorithm, (ii) prepopulation by means of COMBO, a state-of-the-art joint lemmatizer, PoS tagger, and parser trained on the data of the Latin Dependency Treebank 2.1, and (iii) manual error correction informed by the attempt to identify and document lemmatization and morphology annotation rules.

Keywords: Latin Dependency Treebank, lemmatization, PoS tagging.

1. *Introduction*

Lemmatization is, in computational linguistics, a task which is commonly considered part of the morphological layer of annotation because of the strong interrelationship between it and PoS tagging (including morphological features identification)[1], all of them concerning the forms a given word can take on the basis of its function in a clause[2].

[1] A note on the terminology I use in the paper. The expression 'morphological analyzer' is used to mean a program outputting morphological analyses for tokens out of context (e.g. the neuter noun *bellum* receives three analyses, the nominative, accusative, and vocative, which share the same word form). The expression 'morphological analysis' is usually used with reference to a morphological analyzer. On the contrary, I use 'PoS tagger' to mean a statistical tagger outputting one single analysis for each token depending on the context; 'PoS tagging' can however also be used in a general sense, i.e. with or without reference to a PoS tagger. The term 'lemmatizer' is used for programs providing lemmas for tokens both in context and out of context.

[2] Both lemmatization and PoS tagging crucially depend on tokenization, as is shown in Section 4.2.

Received: January 2020
Accepted: May 2020

More precisely, lemmatization can be defined as consisting in the assignment of an 'ID word form' to a set of word forms sharing the same 'base' or 'root' and the same 'part of speech'. An example is the Latin verb *collaboro* ("I collaborate") as the lemma for all the verb word forms sharing the base *collabor*, such as *collaboravit*, *collaboravissemus*, *collaborare*, *collaboratum*, and so forth. Similarly, the Latin noun *dux* is the lemma for all the noun word forms whose base/root is *duc*, such as, for example, *duci*, *ducem*, or *duces*.

Very often, morphologically related word forms share the same 'root', but have different bases. For example, third conjugation verbs, such as *cado*, *is*, *cecidi*, *casum*, *cadere*, typically present different stems. Latin nouns, such as *artifex*, *icis*, can also show vowel changes between the nominative and all other cases.

It is important to note that the choice of a given word form as the ID for its morphologically related word forms is arbitrary/conventional. In Latin, for example, the first person singular of the indicative present is chosen as a lemma (such as *collaboro* above), even though any other related verb word form could in theory be chosen. Indeed, the infinitive form of a verb is commonly used, for example, in Italian, to serve the same ID function as the indicative present in Latin.

It is as important to note that the inventory of parts of speech is also, at least to a certain extent, rather arbitrary/conventional[3]. In Latin, for example, the participle could be considered as an independent part of speech because of its peculiar morphosyntactic proprieties, which, as the etymology of the name itself reveals (*particeps*, i.e. it takes part in the nature of both verb and noun) set it apart from other verb forms.

In lexicography/traditional grammar, lemmas correspond to dictionary entries. Crucially, such entries commonly correspond to more than a single word form: for example, the dictionary entry/lemma for the above mentioned verb *collaboro* is *collaboro*, *collaboras*, *collaboravi*, *collaboratum*, *collaborare*, i.e. it contains, besides the first person singular of the indicative present, also the second person singular of the indicative present, the first person singular of the indicative perfect, the supine, and the infinitive. All these forms provide full morphological information about the verb, because all relevant verb stems are provided, which allow analysis/identification of any word form of the verb.

Dictionary lemmas are most useful because Latin verbs belonging especially to the third conjugation can show unpredictable verb stems: for ex-

[3] How problematic definition of parts of speech is is made particularly clear in typological studies (see, among many others, Haspelmath, 2012 and Sasse, 2001).

ample, the dictionary entry for *capio* ("I take") is *capio*, *capis*, *cepi*, *captum*, *capere*, where the stems for the perfect, supine, and even the infinitive are not as regular as, for example, those of most verbs of the first conjugation. The lemma provided in a dictionary entry is, therefore, aimed not only to function as an ID for the set of its morphologically related word forms, but also to provide full information for its conjugation/declension.

On the contrary, a lemma in treebanking conventionally consists only in the first word form of its corresponding dictionary entry. This has a significant impact on further automatic processing of a given token. For example, the lemmas for the word forms *lupum* and *exercitum* are *lupus* and *exercitus*, respectively. Without knowing their corresponding dictionary lemmas (i.e. *lupus*, *lupi* and *exercitus*, *exercitus*), it is impossible to correctly decline them, even if one takes their morphological analyses into account: indeed, they are both masculine, singular, and accusative nouns.

The information concerning their kind of declension (i.e. I decl. vs IV decl.), which is necessary to correctly decline them, is simply missing in the annotation available within treebanks[4]. This deficiency is even more apparent when it comes to verbs: it is not possible to infer all verb stems from a given word form such as, for example, *ausum* (whose lemma would be *audeo*). Apart from most first conjugation verbs, there is no way to automatically infer all verb stems from single word forms, many of them being potentially able to belong to different conjugations.

As is well known, lemmatization is of crucial importance for many natural language processing tasks, such as summarization, topic modeling, and, more in general, any kind of semantics-oriented research, in that it allows reduction of the variety of word forms available in a text, with consequent increase of machine learning algorithms' performance[5].

In the present article, I will review (some of) the resources available for Latin lemmatization and morphological analysis in Section 2. In Section 3, I draw attention to a few challenges in Latin lemmatization and morphological analysis. In Section 4, I show the current approach to lemmatization and morphological analysis/PoS tagging for the Latin Dependency Treebank. Section 5 contains some concluding remarks.

[4] For an introduction on the treebanks I will mention in the present article, see Celano (2019b) and references therein.

[5] An interesting, recent example of the potential of lemma information for Latin research is presented in Sprugnoli *et al.* (2019).

2. *An overview of Latin lemmatizers and morphological analyzers*

There exist many lemmatizers/morphological analyzers for Latin nowadays, and their number is likely to grow due to the increasing availability of digitized texts/corpora and accessibility of machine learning techniques. I will show in the present section (some of) the most known/remarkable ones[6].

Lemmatizers/morphological analyzers can be evaluated along different dimensions. For example, their coverage of the Latin vocabulary varies. A systematic comparison of all of them is missing, but Springmann *et al.* (2016) provide evidence[7] that *LatMor*[8] (Springmann *et al.*, 2016) and *LemLat*[9] (Passarotti *et al.*, 2017) can recognize many more types/tokens of Classical and Medieval Latin than *PROIEL*[10], Parsley[11], Words[12], and Morpheus[13].

Some lemmatizers/morphological analyzers seem to have been primarily created for human, rather than machine, consumption. For example, both Words and Collatinus[14] can be queried via HTML interfaces or desktop applications, which make them useful especially for traditional scholarship. Words could also be queried automatically because word forms to analyze are contained in URLs[15]: however, the output is a simple HTML page providing no structure for its morphological analyses, so automatic extraction is not immediate. The sources for its more than 39,000 entries seem to derive from the *Oxford Latin Dictionary* and Lewis and Short[16].

Collatinus[17] is based on: Lewis and Short (1879), Gaffiot (2016), Du Cange (1883), Georges (1913-1918), Jeanneau (2017), Gaffiot (1934),

[6] Gleim *et al.* (2019) have recently trained a few PoS taggers and lemmatizers for Latin, using data from *PROIEL* and Capitularia. They run a number of interesting experiments, including testing how well a model can perform on a different kind of corpus.

[7] These results are in line with the ones in Gleim *et al.* (2019: 19).

[8] See *http://www.cis.uni-muenchen.de/~schmid/tools/SFST/*.

[9] I always refer to *LemLat* 3.0: *http://www.lemlat3.eu/*.

[10] See *https://github.com/mlj/proiel-webapp/tree/master/lib/morphology*.

[11] See *https://github.com/goldibex/parsley-core*.

[12] See *http://archives.nd.edu/words.html*.

[13] See *https://github.com/tmallon/morpheus*.

[14] See *https://outils.biblissima.fr/en/collatinus*.

[15] An example for *amoris* is *https://archives.nd.edu/cgi-bin/wordz.pl?keyword=amoris*.

[16] I could not find more precise references for the dictionaries on *https://mk270.github.io/whitakers-words/plan.html* [accessed on 30.11.2019].

[17] The references for the source dictionaries which follow coincide with the bibliographically incomplete ones given on the website *https://outils.biblissima.fr/en/collatinus/#downloads* [accessed on 30.11.2019]. To interpret them, the reader is referred to the weblink, from where the relevant resources can be downloaded.

Calonghi (1898), Valbuena (1819), Quicherat (1836). Its last version (11.2) is claimed to contain more than 80,000 lemmas. Notably, Collatinus[18] also outputs information for syllable length and lemmas are provided in their full form, i.e. in the way they can be found in printed dictionaries (the latter feature is present also in Words). The underlying data is available on GitHub[19], but an API for computer consumption is not provided.

Some lemmatizers/morphological analyzers, such as Morpheus and *LemLat* are especially known for their use in treebanking[20]. Morpheus is the morphological analyzer/lemmatizer used for the Ancient Greek and Latin Dependency Treebank (it will be introduced in Section 4).

LemLat shares the same annotation scheme with the Index Thomisticus Treebank. Even though *LemLat* is one of the oldest lemmatizers/morphological analyzers for Latin, its source code and data have been made open much later[21] (which impacted its exploitation in other projects). It consists in a rule-based morphological analyzer, which depends on a MySQL database containing the data for lemmas/morphological forms.

The internal workings are described in the corresponding documentation[22]. It can be queried within a standalone application, which outputs morphological analyses and lemmas for each word form given as an input. Notably, *LemLat* provides a segmentation for each word analyzed, which distinguishes bases from endings (this information is provided also in Words). The possibility to download the entire database as a MySQL dump guarantees even more query flexibility[23]. The database is based on Georges (1913-1918), Glare (1968-1982), and Gradenwitz (1904), which together amount to 40,014 lemmas, and *Totius Latinitatis Onomasticon* (26,415 lemmas; see Passarotti *et al.*, 2017 for more details).

LatMor is a finite-state morphological analyzer which parses Latin words and returns their morphological analyses, lemmas, and, notably, even vowel quantities. It is accessible at the command line, after the SFST

[18] I refer to the version available online [accessed on 30.11.2019].

[19] See *https://github.com/biblissima/collatinus/tree/master/bin/data*.

[20] A backoff Latin lemmatizer based on the data of the Latin Dependency Treebank is available in *CLTK*: *http://docs.cltk.org/en/latest/latin.html*.

[21] More precisely in 2016, if one follows the date of creation for the corresponding GitHub repository: *https://api.github.com/respos/circse/lemlat3*.

[22] See *https://github.com/CIRCSE/LEMLAT3*.

[23] Because of the complexity of the rules governing the merging of the morphological forms contained in the many MySQL tables, a desideratum for the future is rearranging the content of the database and publish it also in other formats.

tools have been installed. It is based on the lemmas found in Georges (1913-1918) and additions from Lewis and Short (1907); it contains about 70,000 lemmas.

There are a few problems affecting all the above mentioned lemmatizers/morphological analyzers. All of them cannot communicate among each other without proper conversion of morphological labels, since they are all different[24]. The annotation schemes are, in general, similar, but there are still differences, which require attention. For example, *ubi* is classified as 'invariable' in *LemLat*, but as 'adverb' or 'conjunction' in *LatMor*.

Another remarkable problem is that each lemmatizer/morphological analyzer joins together lemmata of more than one dictionary on the assumption that there is consistency across all the resources as to the criteria employed to identify lemmata. This probably holds true in general (also because of the known interdependencies among the original printed editions), but it is still unknown to what extent exactly.

One technical limitation of all the lemmatizers/morphological analyzers is that they cannot analyze multiword expressions, such as passive forms: for example, the expression *amatus est* cannot be given as an input and analyzed as a perfect passive indicative, but it has to be split into *amatus* (i.e. 'perfect passive participle') and *est* ('present indicative'). This is unfortunately an unsolved problem also affecting treebanking, where tokenization typically allows splitting but not merging of two graphic words, and therefore multiword tokens such as *amatus est* can be annotated only by means of specific syntactic labels[25].

Lastly, none of the lemmatizers/morphological analyzers provides a community-based mechanism allowing editing of the databases, which could guarantee corrections and extension. Most resources do not make the underlying database open or easily accessible; when the database is available (such as those of *LemLat* or Collatinus), their formats do not allow editing easily.

[24] Differences in orthography may also apply.

[25] The inability of properly analyzing multiword expressions in treebanking heavily depends on the fact that tokenization and morphosyntactic annotation are not commonly added standoff: inline annotation makes it difficult to express splitting and merging of graphic words, while trying to keep markup in a file relatively simple and easy to understand (and process). For a proposal of standoff annotation for Latin see Celano (2019a).

3. *Challenges for (Latin) lemmatization and morphological analysis*

There exist challenges concerning lemmatization and morphological analysis for Latin (as well as other languages), which especially pertain to the computational nature of these tasks.

As was shown in Section 2, most lemmatizers/morphological analyzers rely on information contained in more than one printed dictionary. This raises the question of which criteria were employed (i) to identify lemmas and – for the purposes of morphological analysis – (ii) to assign them a part of speech.

These two problems particularly affect digital resources because they should strive to ensure as much consistency as possible, any automatic data processing crucially relying on it. Indeed, while consistency is also desirable in printed dictionaries, it seems reasonable to claim that the specific purpose for which they were created (i.e. human consumption) may allow for accommodation of a number of 'irregularities', which, on the contrary, impinge on computational resources derived from them, but designed for machine consumption.

This is particularly clear when it comes to deciding about the part of speech for a given word: for example, *hiberna* could be analyzed as a substantivized adjective and therefore subsumed under the adjectival lemma *hibernus*, *a*, *um* or considered as a separate noun, and therefore assigned the separate entry *hiberna*, *orum*. Some printed dictionaries opt for the second solution, but it is not completely clear why: on the one hand, *hiberna* seems to occur so frequently as to be able to be recognized as belonging to an independent – although related – lemma; on the other, neuter adjectives can be regularly substantivized in Latin, but many/most of them are subsumed under their corresponding adjectival lemmas (similarly, *Romani* is, for example, found under *Romanus*, *a*, *um*).

Strictly connected to the question of substantivized adjectives is that of participles. Participles are regularly assigned the part of speech 'verb', even though, as is well known, they can serve different functions within a clause. Classification of participles in printed dictionaries is different and not even always consistent within the same dictionary.

For example, the *Oxford Latin Dictionary* (Glare, 1968-1982) has two different lemmas for *amans*, the former being an adjective and the latter a noun. However, *subiectus* is there presented only as an adjective lemma, with

its function as a noun being a subcategorization of it. On the contrary, Georges (1913-1918), has only one lemma for *amans* (as an adjective and a noun), but two for *subiectus* (the adjective function being kept separate from the noun one). Both dictionaries, however, do not consider *laborans* a lemma, even though it is also attested with the meaning of "the one who works".

In *LemLat* ('BASE LES' function) *amans* is analyzed as a noun (not as an adjective), while *subiectus* (i.e. "a subordinate") as a verb. On the contrary, *LatMor* keeps the adjective and the noun lemmas separate both for *amans* and *subiectus*. Another interesting example is *florens*, which is analyzed as a verb (i.e. participle) in *LemLat*, but as an adjective and a verb in *LatMor*.

A classification issue similar to that of participles is posed by infinitives. They are normally analyzed as verbs, but one should note that infinitives functioning as nouns are also classified as verbs and therefore their lemmas correspond with that of the corresponding verbs: this clearly challenges the annotation scheme's consistency/uniformity, in that the category 'noun', which is acknowledged, for example, for *studium*, should/could in principle also apply, for example, to *studere* in *studere bonum est*.

Likewise, the gerundive and gerund are problematic because of their nature at the interface between 'verb' and, respectively, 'adjective' and 'noun'. This becomes evident at the syntactic level, in that it is questionable whether they should get adjectival/nominal or verbal syntactic labels.

Rather idiosyncratic is also the category 'pronoun', which does not distinguish pronouns used as adjectives (e.g. *horum amicorum*) from those used as nouns (e.g. *horum*). PoS tagging for relative adverbs such as *ubi*, *quo*, and *qua* can fluctuate between 'adverb' and 'conjunction': in *LemLat ubi* is 'invariable', while *quo* and *qua* are 'pronominal'; in *LatMor ubi* and *quo* are both 'adverb' and 'conjunctions', but *qua* is only 'adverb'.

All the above mentioned uncertainties arising in lemmatization/morphological analysis are ultimately due to lack of (clear) definitions for morphological categories. This is a long-standing problem in linguistics. However, while such classification inconsistencies in printed dictionaries can usually be accommodated by readers because lemmas and the corresponding PoS labels primarily serve the purpose of pointers to word meanings, they impact lemmatizers/morphological analyzers much more severely, in that their function is supposed to be that of providing reliable lemmatization/morphological classification.

Moreover, printed dictionaries can much better cope with spelling issues. It is well known that over centuries Latin showed spelling variants, which lexicographers often try to account for by using internal references. If one looks up *adpono* in, for example, the *Oxford Latin Dictionary* (Glare, 1968-1982), a reference to *app-* is given the reader for all words starting with *adp-*. This system is also used for 'grammatical' references: in the *Thesaurus Linguae Latinae*, for example, *florens* refers to the verb *floreo*.

LemLat has an internal converter for spelling variations: for example, it automatically converts *v* into *u*. This feature is not present in *LatMor*: if a word has a different spelling, it is simply not recognized. Contrary to *LemLat*, *LatMor* adopts the distinction between consonantal *u* (spelled as *v*) and vocalic *u*.

4. *The Latin Dependency Treebank: Towards guidelines for morphological annotation*

Strange as it may sound, there are no guidelines for morphological annotation for the Latin Dependency Treebank (as well as for the other Latin treebanks). Annotation of morphology may at first sight seem less difficult/problematic than that of syntax, and admittedly many studies have been produced for Latin morphology over the centuries, which have reached a consensus on many key points.

Notwithstanding, accounts for Latin morphology vary and, as was shown in Section 3, there exist a number of open questions that need to be addressed before performing corpus annotation. In the light of that, the Latin Dependency Treeebank is currently under revision[26]: in the present section, I outline the current pipeline to annotate lemmas/morphology in it and the challenges faced to foster consistency.

I will discuss the problem of orthography and tokenization in Section 4.1 and Section 4.2, respectively. I will present the morphological analyzer Morpheus in Section 4.3 and the COMBO lemmatizer/PoS tagger/parser in Section 4.4.

[26] *DFG* project 408121292: *https://gepris.dfg.de/gepris/projekt/408121292?context=projekt&task=showDetail&id=408121292&*.

4.1. *Orthography*

An underestimated annotation problem is that of orthography. Latin, as is known, has been written differently over the centuries, and such variations are sometimes recorded in critical editions.

Among the most well-known variations is that between the letters *u* and *v*, and *i* and *j*. Classical Latin had only one letter for both /ʊ uː/ and /w/ and one letter for both /ɪ iː/ and /j/. The two oppositions between the consonant and vowel sounds were introduced in writing later. Other well-known variants – just to mention a few – are the groups *adp-*/*add-*, *adn*/*ann-*, or vocalic alternations such as that in *seruos*/*seruus*.

Such variants pose a challenge for text digitization, lemmatizers/morphological analyzers, and PoS taggers. In the Latin Dependency Treebank, digitized texts preserve the Latin spelling found in critical editions. A normalization layer is, however, planned to be added standoff to each text, so that texts with different spellings can be queried easily and efficiently.

The normalization layer relies on Brambach's rules (McGabe, 1877)[27], which promote use of Latin orthography of the Silver Age. In offering clear guidelines, Brambach's system has already been adopted by many editors[28].

4.2. *Tokenization (and sentence split)*

Tokenization[29] consists in identifying the minimal units for a given analysis/annotation. It is fair to say that the tokenization task for Latin has received much less attention than it deserves. Tokenization represents, *stricto sensu*, the first kind of annotation a text receives.

It is not clear how to exactly define what a token should be in morphosyntactic analysis[30]. In Latin, for example, the negation *non* and the conjunction *et* are recognized as (separate) tokens, but in some treebanks *nec* (i.e. *et non*) represents a single token. Another example are multiword

[27] See *https://archive.org/details/aidstolatinortho00bramrich/page/n6*.

[28] Notably, Brambach sometimes offers more than one option. For more information on how these cases are dealt with, see *https://git.informatik.uni-leipzig.de/celano/latinnlp/blob/master/guidelines/01_orthography.md*.

[29] 'Tokenization' is here used to describe the processes that are sometimes referred to by some scholars as 'tokenization' and 'word segmentation'.

[30] This is of course related to the well-known open question of definition of 'word' (see, for example, Simone, 2008: 150 ff. for a few examples of its heterogeneous nature).

expressions, such as *res publica*: they are commonly treated as two tokens, even though they function syntactically as one-word tokens, such as *Roma* or *mare*.

The problem seems to be even more challenging when it comes to finding a definition of token that applies crosslinguistically: the function of prepositions in a language can be, for example, expressed by cases in another language. One attempt to mitigate some of the irregularities of current tokenization schemes is to account for them at the syntactic level via the use of specific syntactic labels.

Clearly, such a strategy, which seems to be dictated by convenience[31], is questionable on a theoretical level. It is also untested what the impact of such a strategy is on, for example, PoS taggers/syntactic parsers.

For the Latin Dependency Treebank a new rule-based algorithm[32] has been developed to tokenize texts. After whitespace-based tokenization, if a token ending with a punctuation mark is not recognized as an abbreviation (via the use of a word list and a regular expression), the punctuation mark is separated. The same token is then analyzed to see if it matches one of the members in a list containing tokens which need to be split by *ad-hoc* rules: this holds true, for example, for *mecum* or *nequis*.

In order to avoid inconsistencies in the treatment of expressions such as *postquam* and *post quam* or *etiamnunc* and *etiam nunc*, the above mentioned list also contains those tokens that are recognized to have the same function/meaning but can be written as one or two tokens[33]. The split is preferred over the univerbated variant for two reasons: the split variant (i) (typically) antedates the univerbated form and (ii) it is easier to formalize splitting than merging, in that the parts of a split token such as *postquam* could not be adjacent in a clause.

Finally, a graphic word is split into two tokens if it contains the enclitics *que*, *ve/ue*, and *ne*, including *neque*, *nec*, *neve*, *neue*, and *neu*. These latter were sometimes treated as single tokens in the past. They are however split today according to the principle whereby a token needs to be identified if it is required in order to build a correct syntactic tree. For example, if *neque* were not split, one could not correctly build the tree for a sentence

[31] Tokenization asymmetries seem to be related to lack of standoff annotation.

[32] For full documentation, including the actual algorithm, see *https://git.informatik.uni-leipzig.de/celano/latinnlp/blob/master/guidelines/02_tokenization.md*.

[33] This holds true especially for texts of the Golden/Silver Age, which are currently the focus of the Latin Dependency Treebank.

such as the following (I have abbreviated the sentence to focus on the issue at hand):

(1) *Omnes Belgarum copias* [...] *ad se venire vidit neque iam longe abesse* [...] *cognovit.*
"He saw all troops of the Belgae [...] were approaching toward him and learned that they [...] were then not far distant."[34]
(Caes. *De Bello Gallico* 2.5.4)

The conjunction *que* coordinates *vidit* and *cognovit*, but the negation *ne-* applies to *abesse* (and not to *cognovit*).

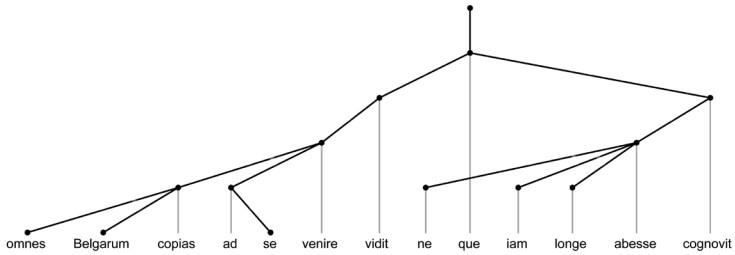

Figure 1. *Parse tree for Caes.* De Bello Gallico *2.5.4.*

Like tokenization, sentence split is currently performed rule-based via a simple algorithm which identifies the major punctuation marks, i.e. full stop, colon, semicolon, question mark, and exclamation point[35].

4.3. *The morphological analyzer Morpheus*

Morpheus (Crane, 1991) is available on the Perseus website[36], via a web API[37], and as a MySQL dump downloadable from the Perseus website[38] (it is also integrated into the annotation tool Arethusa)[39]. These instances serve different purposes. The Perseus website allows easy human interrogation, with morphological analyses been also connected to other resources such as the Lewis and Short (1879) dictionary.

[34] The translation follows *http://data.perseus.org/texts/urn:cts:latinLit:phi0448.phi001.perseus-eng1*.

[35] See *https://git.informatik.uni-leipzig.de/celano/latinnlp/blob/master/guidelines/04_sentence_split.md*.

[36] See *https://www.perseus.tufts.edu/hopper/morph?l=amoris&la=la*.

[37] See *https://www.perseus.tufts.edu/hopper/xmlmorph?lang=lat&lookup=cepissem*.

[38] See *https://www.perseus.tufts.edu/hopper/opensource/download*.

[39] See *https://sosol.perseids.org*.

The web API is designed to automatically parse Latin word forms. The API returns an XML document containing as many <analysis/> elements as the number of possible analyses for a given word form. For example, two analyses for *donis* are given, in that this word form corresponds to the dative plural and ablative plural of the lemma *donum*.

Each <analysis/> element contains a number of child elements describing the morphology of the word form. Among these are the <lemma/> element and <pos/> element (i.e. part of speech), as well as other elements describing morphological features, such as <number/> and <gender/>.

It is possible that the above mentioned versions slightly vary from each other. In the MySQL dump the hib_lemmas table contains 17,573 Latin lemmas. The hib_parses table contains possible morphological forms for each lemma in the hib_lemmas table (466,748). Joining the two tables via the lemma_id field easily allows getting all the word forms and their analyses for a given lemma.

Latin Morpheus is based on the Lewis and Short (1879) dictionary entries. The format of its morphological analyses coincides with the one used in the Latin Dependency Treebank. It is therefore used, for example, to suggest possible morphological analyses during annotation in Arethusa.

The annotation scheme for morphology consists in a 9-character long string, each of them always corresponding to a specific morphological category, which can take one of a finite set of values: if a given category does not apply to a word form, a hyphen is used. The first character specifies the part of speech, and can be any of the following: noun, verb, (participle), adjective, adverb, conjunction, preposition, pronoun, numeral, interjection, and punctuation.

In Morpheus it is possible to see participles treated as a part of speech, but in the Latin Dependency Treebank, 'participle' is a mood. The remaining eight characters represent the following morphological categories[40]: person, number, tense, mood, voice, gender, case, and degree. For example, *rumores* can be annotated as 'n-p---ma-', i.e. noun plural masculine accusative.

As showed previously, there are a few issues concerning morphological annotation and lemmatization that require guidelines. For the next release of the Latin Dependency Treebank, all substantivized nouns are lemma-

[40] See for the sets of all values *https://git.informatik.uni-leipzig.de/celano/latinnlp/blob/master/guidelines/03_morphology.md*.

tized under the corresponding adjective lemmas. This is done in that common practice has always been to generally not identify new lemmas for substantivized adjectives (see, for example, *Romani* as "the Romans", which is typically found under *Romanus*, *a*, *um*). This choice is made also because it seems to be in agreement with the treatment of similar phenomena: for example, substantivized participles are also commonly lemmatized under the corresponding verbs, and pronouns are also not distinguished in their adjectival and nominal function.

Relative adverbs, such as *ubi*, *quo*, or *qua* should be tagged as adverbs, even when they are used without an antecedent and their function resembles that of a conjunction. Indeed, the risk in treating them as conjunctions is that, if any of them happens to play the role of an argument, this can correctly be annotated only if the token is tagged as an adverb.

It is probably because of argument structure that sometimes *ubi* meaning "when" is classified as 'conjunction', while *ubi* meaning "where" tends to be considered as a 'relative adverb': the former is typically an adjunct. Similarly, *quo* meaning "to where" is typically an argument and therefore tends to be analyzed as a relative adverb.

Because of the great variety of lemmatization peculiarities which can affect single tokens and because of the fact that dictionaries are not always consistent in and among themselves as to lemmatization/PoS tagging, the best approach in creating digital resources is probably to make available, and regularly update, open lexica (both for human and computer consumption) compiled following documented criteria.

4.4. *The COMBO lemmatizer/PoS tagger/parser*

Currently, texts in the Latin Dependency Treebank are prepopulated both for lemmatization/morphology and syntax using the output of COMBO. After that, they are typically ingested in the Arethusa annotation tool, so that errors can be manually corrected.

COMBO (Rybak and Wróblewska, 2018)[41] is a state-of-the-art joint neural lemmatizer, PoS tagger, and parser which ranked among the best ones in the *CoNLL* 2018 Shared Task. More precisely, it ranked as the 4th best parser for UPoS, 5th for XPoS, 3rd for morphological features, and 7th for all morphological tags (all rank positions concern annotation of the

[41] See *https://github.com/360er0/COMBO*.

Latin data of the *UD* Latin Dependency Treebank). Differently from other parsers, COMBO has been made available online and is relatively easy to retrain.

As the *CoNLL* 2018 Shared Task is based on data annotated in the Universal Dependency annotation scheme, COMBO had to be retrained in order to output annotations according to the annotation scheme of the Latin Dependency Treebank (v. 2.1). Table 1 shows the accuracies for lemmatization and PoS tagging; the models, a REST API, and accuracies for the syntactic annotation are available online[42].

Field	*Accuracy*
LEMMA	0.83
PoS	0.90
XPoS	0.72
FEAT	0.74

Table 1. *Accuracies for Latin*.

The REST API provided for COMBO allows outputting of morphological and syntactic annotation for Latin according to different annotation schemes: Latin Dependency Treebank, *UD* Latin Dependency Treebank, *UD* Index Thomisticus Treebank, *UD PROIEL* Treebank (the *UD* models are available on the COMBO GitHub repository).

5. *Conclusion and prospects*

The present paper has presented some challenges posed by lemmatization and morphological analysis for Latin, with reference to the ongoing work for the revision of the Latin Dependency Treebank. It has been argued that lemmatizers/morphological analyzers mostly depend on digitized dictionaries, which however contain a number of inconsistencies in lemma identification and PoS tagging.

Indeed, printed dictionaries have been created primarily to provide definitions for Latin words, rather than consistent lemmatization. On the contrary, digital resources, such as treebanks, need to aim to classify Latin

[42] See *https://git.informatik.uni-leipzig.de/celano/COMBO_for_Latin*.

tokens as consistently as possible in order to facilitate automation and query of annotations.

Annotation for the Latin Dependency Treebank currently relies on a rule-based tokenization and sentence-split algorithm, whose output feeds the COMBO lemmatizer, PoS tagger, and parser, used to prepopulate texts. Subsequently, both lemmas and morphological labels are manually corrected. Within the Arethusa annotation tool, the morphological analyzer Morpheus can sometimes help selection of correct alternative labels.

A major goal of the current revision of the Latin Dependency Treebank is to also document annotation choices for lemmatization/morphology via examples/rules to foster consistency: this is work in progress[43].

Acknowledgments

I gratefully acknowledge the support of the *DFG* (Deutsche Forschungsgemeinschaft), which has funded the present research (Project Number: 408121292).

References

Celano, G.G.A. (2019a), *Standoff annotation for the Ancient Greek and Latin Dependency Treebank*, in *DATeCH2019: Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage, Brussels, Belgium*, Association for Computing Machinery, New York, pp. 149-153.

Celano, G.G.A. (2019b), *The Dependency Treebank for Ancient Greek and Latin*, in Berti, M. (2019, ed.), *Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution*, De Gruyter, Berlin, pp. 279-298.

Crane, G. (1991), *Generating and parsing Ancient Greek*, in «Literary and Linguistic Computing», 6, 4, pp. 243-245.

Georges, K.E. and Georges, H. (1913-1918), *Ausführliches Lateinisch-Deutsches Handwörterbuch*, Hahn, Hannover.

Glare, P.G.W. (1968-1982), *Oxford Latin Dictionary*, Oxford University Press, Oxford.

[43] See *https://git.informatik.uni-leipzig.de/celano/latinnlp*.

Gleim, R., Eger, S., Mehler, A., Uslu, T., Hemati, W., Lücking, A., Henlein, A., Kahlsdorf, S. and Hoenen, A. (2019), *Practitioner's view: A comparison and a survey of lemmatization and morphological tagging in German and Latin*, in «Journal of Language Modelling», 7, 1, pp. 1-52.

Gradenwitz, O. (1904), *Laterculi Vocum Latinarum*, Hirzel, Leipzig.

Haspelmath, M. (2012), *How to compare major word-classes across the world's languages*, in «UCLA Working Papers in Linguistics», 17, pp. 109-130.

Lewis, C.T. and Short, C. (1879), *A Latin Dictionary*, Oxford University Press, Oxford.

Lewis, C.T. and Short, C. (1907), *A New Latin Dictionary*, American Book Co., New York.

McGabe, W.G. (1877), *Aids to Latin Orthography by Wilhelm Brambach*, Harper and Brothers, New York.

Passarotti, M., Budassi, M., Litta, E. and Ruffolo, P. (2017), *The* Lemlat *3.0 package for morphological analysis of Latin*, in Bouma, G. and Adesam, Y. (2017, eds.), *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language, Gothenburg, Sweden*, Linköping University Electronic Press, Linköping, pp. 24-31.

Rybak, P. and Wróblewska, A. (2018), *Semi-supervised neural system for tagging, parsing and lemmatization*, in Zeman, D. and Hajič, J. (2018, eds.), *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Brussels, Belgium, October 31 - November 1*, Association for Computational Linguistics, pp. 45-54.

Sasse, H.-J. (2001), *Scales between nouniness and verbiness*, in Haspelmath, M., König, E., Oesterreicher, W. and Raible, W. (2001, eds.), *Language Typology and Language Universals*, De Gruyter, Berlin / New York, pp. 495-509.

Simone, R. (2008), *Fondamenti di linguistica*, Laterza, Roma / Bari.

Springmann, U., Schmid, H. and Najock, D. (2016), LatMor*: A Latin finite-state morphology encoding vowel quantity*, in Celano, G.G.A. and Crane, G. (2016, eds.), *Treebanking and Ancient Languages: Current and Prospective Research*, in «Open Linguistics», 2, 1, pp. 386-392.

Sprugnoli, R., Passarotti, M. and Moretti, G. (2019), Vir *is to* Moderatus *as* Mulier *is to* Intemperans *- Lemma embeddings for Latin*, in Bernardi, R., Navigli, R. and Semeraro, G. (2019, eds.), *Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15*.

Giuseppe G.A. Celano
Abteilung Automatische Sprachverarbeitung
Institut für Informatik
Universität Leipzig
Augustusplatz 10
04109 Leipzig (Germany)
celano@informatik.uni-leipzig.de

SSL

CLaSSES: Orthographic variation in non-literary Latin

Giovanna Marotta, Francesco Rovai,
Irene De Felice, Lucia Tamponi

Abstract

CLaSSES (Corpus for Latin Sociolinguistic Studies on Epigraphic textS) is a digital resource which gathers non-literary Latin texts (epigraphs, writing tablets, letters) of different periods and provinces of the Roman Empire. This corpus has been tagged with linguistic and extra-linguistic information that allows quantitative and qualitative analysis of spelling variations in Latin sources. The resource is available on the web in open access and is structured in different sections: *Rome and Italy*, *Roman Britain*, *Egypt and Eastern Mediterranean*, and *Sardinia*.

Keywords: Latin digital resources, Latin non-literary texts, historical sociolinguistics.

1. *Introduction*

CLaSSES, i.e. Corpus for Latin Sociolinguistic Studies on Epigraphic textS, is a digital resource that contains non-literary Latin texts (epigraphs, writing tablets, documentary letters) of different periods and provinces of the Roman Empire. The database is available on the web in open access (*http://classes-latin-linguistics.fileli.unipi.it*) and has been developed at the Laboratory of Phonetics and Phonology of the Department of Philology, Literature and Linguistics at Pisa University[1].

This new resource joins a growing list of digital tools (epigraphic collections, lemmatizers, syntactic treebanks, etc.) suitable for academic research on the Latin language, of which a representative sample is provided in this issue of *Studi e Saggi Linguistici*. In-between the lemmatizers and

[1] The construction of the corpus began during the PRIN project *Linguistic representations of identity. Sociolinguistic models and historical linguistics* (PRIN 2010, prot. 2010HXPFF2_001). The initial plan of the database included only the section *Rome and Italy*. Over the last few years, the sections *Roman Britain*, *Egypt and Eastern Mediterranean*, and *Sardinia* have been added, while maintaining the original structure and layout (see below, § 3).

Received: February 2020
Accepted: May 2020

treebanks, whose reference corpora are mostly based on the literary texts, and the available digital epigraphic collections, which are not specifically designed for linguistic research, *CLaSSES* is a representative corpus of epigraphic and other non-literary documents annotated with linguistic information.

The working hypothesis is that non-literary texts (inscriptions, ostraka, documentary papyri, private letters, ink tablets) can be a direct and reliable source in order to approach a picture of the sociolinguistic variation that characterized the Latin-speaking world. In particular it is the (ortho-)graphic variants, as testified by the misspellings (i.e. those spellings that are not congruent with the 'standard' language as exhibited in the literary texts of the Classical period) occurring in non-literary texts, which can be assumed as being clues for linguistic variation.

Of course the current debate on the reliability of the inscriptional evidence for the investigation of linguistic variation and change in the ancient languages, is polarized between more or less skeptical views (cf. e.g. Adams, 2007; 2013 vs Herman, 1985). On the one hand, inscriptions, ink tablets, ostraka, and papyri are the only direct, first-hand evidence left from antiquity, while in any other kind of written text the mediation of the later philological and manuscript tradition is present. On the other hand, their value has always to be checked against a fine-grained analysis of the philological, paleographic, archaeological, and historical aspects, in order to reduce the problem of data sparseness that originates from the fragmentary nature of non-literary texts, as well as the problem of the authorship of the text.

At present, several scholars believe that non-literary texts can be regarded as a fundamental source for studying language variation (e.g. Molinelli, 2006; Kruschwitz, 2015; Marotta, 2015; Rovai, 2015; Consani, 2016), so that studies on the sociolinguistic aspects of Latin in Rome and the Empire have recently flourished[2], although some seminal works date back to some decades ago (e.g. Campanile, 1971; Vineis, 1984; 1993).

Building on this hypothesis, *CLaSSES* has been specifically designed in order to collect non-literary documents which attest spelling variants that could be indicative of phenomena occurring in the phonological or morpho-phonological realms. Such orthographic variants have been labelled as

[2] See for instance Adams (2003; 2007; 2013); Rochette (1997); Biville *et al.* (2008, *éds.*); Dickey and Chahoud (2010, eds.).

'non-classical' forms, with reference to the standard spelling forms of Classical Latin. For every non-classical form, the corresponding classical one is also presented. For instance, a form like <MENOS> has been considered non-classical, since its corresponding form in the classical orthographic norm is <MINVS>.

Before illustrating the structure of this paper, an introductory methodological consideration is necessary here. Since the database is intended as an instrument for future research, the linguistic annotation of the misspellings is always kept as descriptive as possible and makes reference exclusively to the (ortho-)graphic level. Thus, the phenomena annotated in the case of <MENOS> for <MINUS> are labelled as *Vowel alternation - Classical* <I>, /ĭ/ = <E> and *Vowel alternation - Classical* <U>, /ŭ/ = <O>. This is tantamount to saying that a short *i*-sound of the Classical Latin is represented here through the letter <E> and that a short *u*-sound of the Classical Latin is represented through the letter <O> – and both spellings are inconsistent with the standard Classical orthography, where <I> and <U> occur. No preliminary assumption is therefore made about a possible relative chronology of the two variants, neither in the light of the etymological criterion nor in view of otherwise well-attested patterns of phonetic change. Whether these phenomena can be regarded as the relics of older spellings, or as an early anticipation of a Proto-Romance development, is left to the researcher's conclusive subsequent interpretation.

This paper is structured as follows: Section 2 provides a short review of the digital resources currently available for Latin epigraphy; in Section 3 the documents contained in *CLaSSES* are described with reference to the kind of material and the area of provenance; Section 4 presents the criteria of annotation by which textual data have been implemented with linguistic, meta-linguistic, and extra-linguistic information; Section 5 illustrates the technical aspects for the use of the search interface; finally, in Section 6 we summarize our conclusions.

2. *Digital resources for Latin inscriptions and other non-literary texts: An overview*

In this section, we shortly present the main digital resources available for the study of Latin epigraphy, with reference to database organization, data structure, and the user interface.

2.1. *Epigraphik Datenbank Clauss-Slaby*

Several open-access databases are available online for the study of Latin epigraphy (Feraudi-Gruénais, 2010; Elliott, 2015; cf. also the section *Inschriften in der digitalen Welt* in Eck and Funke, 2014, *Hrsg.*: 501-517) of which the Epigraphik Datenbank Clauss-Slaby (*EDCS*)[3] is at present the most complete digital collection of searchable Latin inscriptions. It records 520,061 texts from 22,232 findspots that cover the entire area of the Roman provinces. Each text is identified with an *EDCS-ID* number and annotated with information containing relevant bibliography, province and findspot. In many cases, further extra-linguistic and meta-linguistic data are provided: dating (for 179,365 inscriptions), material specification (for 187,543 inscriptions), the social status of the people mentioned in the text, and the textual typology (the classification for personal status and inscription genus is conflated under a single heading and is available for 210,844 inscriptions). There are also links (847,661) to other 36 databases, frequently with photos (for 191,027 inscriptions). In order to keep the presentation of the texts as simple as possible, the texts are presented without abbreviations and completed (where possible). The search engine allows for simple and combined word queries also by using Boolean operators and regular expressions, and searches can be limited using various entries of the metadata: records, province, place, dating, material, text type, personal status. Though not specifically designed for linguistic studies, *EDCS* is one of the most valuable sources for the investigation of language variation, since it is possible to search for misspellings (such as CONSENTIONT and TEMPESTATEBUS for *consentiunt* "agree.IND.PRES.3PL" and *Tempestatibus* "Goddess of Storm.DAT.PL.F") through the 'Search entries: wrong spelling' function.

However, it has to be noted that single linguistic forms (either words or groups of letters) rather than linguistic phenomena can be searched and browsed through this function, so that the researcher must already know which form to search for. In this way, as with the other databases described below, there is a risk of sliding into those limitations highlighted by Cordell (2015: 421): «most digital archives hide more than they reveal, as keyword searches require prior knowledge of the texts to be discovered and can lead to evidentiary excess».

[3] Cf. *http://www.manfredclauss.de/* [accessed on 20.02.2020].

2.2. *The* EAGLE *network*

In addition to *EDCS*, reference is to be made to the *EAGLE* Project (Orlandi, 2017; Orlandi *et al.*, 2017, eds.; Prandoni *et al.*, 2017)[4], which began in 2003 as a network of four epigraphic digital archives (Epigraphische Datenbank Heidelberg-*EDH*, Epigraphic Database Roma-*EDR*, Epigraphic Database Bari-*EDB*, and Hispania Epigraphica Online-*HE*) with the aim of assembling the epigraphic collections held by the *EAGLE* partners, in order to provide scholars with a single portal to the searchable inscriptions of the Ancient World. The original four major databases remain pillars of the *EAGLE* network, but an up-to-date overview of the collections represented is available on the website (*https://www.eagle-network.eu/eagle-project/collections/*)[5].

The Epigraphische Datenbank Heidelberg (*EDH*)[6] contains the texts of Latin and Latin-Greek bilingual inscriptions from the Roman provinces, excluding Italy with Sicily, Sardinia and Corsica (for which see Epigraphic Database Roma-*EDR* below), and Spain (for which see Hispania Epigraphica Online below). *EDH* is made up of four constituent parts: Epigraphic Text Database (80,870 inscriptions), Photographic Database (39,031 photos), Bibliographic Database (16,481 records concerning monographs, articles in journals, and other specialist literature), Geographic Database (the geographical details of the 30,272 findspots of the inscriptions included in *EDH*). Users can perform simple full-text searches of words or groups of letters as well as more advanced queries while taking into account the metadata that enrich every single text: findspot, present location, dating (when available), type of inscription (e.g. honorific inscription, epitaph, votive inscription, etc.), language, material (e.g. marble, copper, amber), size and type of the monument (e.g. altar, *cippus*, *sarcophagus*, etc.), writing technique (e.g. engraved, painted, scratched, etc.), and historical relevant data (e.g. religion to which the monument belongs, troop names, people mentioned and rel-

[4] In origin, the acronym was for Electronic Archive of Greek and Latin Epigrahy, but it is now expanded Europeana network of Ancient Greek and Latin Epigraphy.

[5] An important role in the integration of different databases is played by Trismegistos (*TM*; cf. *https://www.trismegistos.org*), which is a central database of metadata (not texts) for papyrological and epigraphic documents from the Greco-Roman world, with a particular focus on prosopographical (*TM* People) and place (*TM* Places) identifications. *TM* currently includes more than 720,000 entries. Since networks of databases such as *EAGLE* and papyri.info (see below, § 2.4) inevitably show duplicate entries for some documents, by using the unique catalog numbers from *TM* (the so-called 'stable identifiers') as identification numbers, users can collate duplicate entries.

[6] Cf. *https://edh-www.adw.uni-heidelberg.de/home* [accessed on 20.02.2020].

ative status, when available, etc.). Each text is also annotated with relevant bibliography and commentary.

The systematic gathering of the inscriptions from Italy and its islands, excluding Christian texts (for which see Epigraphic Database Bari-*EDB* below), is the domain of the Epigraphic Database Rome (*EDR*; Panciera, 2013; Caldelli *et al*., 2014)[7]. Up to date, the *EDR* collection includes 91,336 inscriptions and 59,097 photos. Every text in this database is richly annotated with metadata concerning its dating, findspot and storage place, type of object, material, state and dimension of the support, writing technique of the inscription, language, and text type; when available, personal status of those mentioned in the text is specified. Finally, information concerning relevant bibliography is included. The online query interface allows words or groups of letters to be entered (possibly with Boolean operators) and simple and advanced queries can be made in combination with the following fields: record number, place of provenance (ancient region, current region, ancient city, modern city), current location, object type, material, measurements, state of textual preservation, writing technique, language, religion, verse, inscription type, type of persons mentioned, apparatus, and dating.

Epigraphic Database Bari (*EDB*; Rocco, 2017)[8] is specialized in Christian epigraphic documents from Late Ancient Rome (3rd-7th century AD) and includes 41,602 items and 7,891 images. In addition to the text, for each inscription the following metadata are recorded and featured for the interrogation of the database: graphical (reuse / opisthographic inscription, Greek alphabet), meta-linguistic (metrical text, function), and linguistic (Latin or Greek language) elements, material and executing technique, findspot and current location, dating, and figurative apparatus (*signa Christi*, symbols, various representations). It is of particular interest for linguistic analysis that various options for textual research are featured, including a thesaurus that is intended to search also for misspellings and aberrant forms.

Hispania Epigraphica Online (*HE*)[9] focuses on the epigraphic documents of Portugal and Spain, in large part written in Latin, but with a few examples of Greek, Semitic, and Iberian inscriptions. The corpus includes 30,809 inscriptions, most of which include photos. However, metadata sets of the texts, their degree of elaboration, and search options are less accurate

[7] Cf. *http://www.edr-edr.it/default/index.php* [accessed on 20.02.2020].
[8] Cf. *http://www.edb.uniba.it/* [accessed on 20.02.2020].
[9] Cf. *http://eda-bea.es/* [accessed on 20.02.2020].

than in the above-mentioned databases, so that the search interface holds the following fields: record number, title, object type, inscription type, keyword, inscription, place of finding, place of conservation, and museum.

On the whole, the *EAGLE* network profiles as a massive epigraphic digital resource, which is based on the Metadata Aggregation System (Mannocci *et al.*, 2017: 173-174), i.e. an Aggregative Data Infrastructure (Amato *et al.*, 2013) where all the information of the four major collections illustrated above is stored and indexed. Users can browse the content and interact with it by means of an interface (Prandoni *et al.*, 2014) allowing the searching and browsing of the rich set of data made available by *EAGLE* partners by using either a free text simple search or an advanced search where the user can specify the values of a number of fields. Images can be retrieved through an image recognition algorithm, and translations of the epigraphic texts are available. Finally, it is also possible to export the *EpiDoc* document describing the object[10].

2.3. *Towards a digital epigraphy designed for linguistic research*

As shown in §§ 2.1-2.2, a wide range of digital repositories of epigraphic content are currently accessible online, featuring a great variety of Latin inscriptions, and providing scholars with a cluster of extra-linguistic data, such as provenance place, dating, material, etc. by which to verify the reliability of historical reconstructions. An accurate reconstruction of the socio-historical context is – of course – of primary interest also for the study of language variation and change in the Latin epigraphic (and, more generally, non-literary) documents. In the last few decades, the widely acknowledged dimensions of sociolinguistic variation have proven to be a fertile field of investigation, giving rise to the field of historical sociolinguistics, whose aim is «the reconstruction of the history of a given language in its socio-cultural context» (Conde-Silvestre and Hernández-Campoy, 2012: 1). In particular, many scholars have shown that it is possible to identify different varieties of

[10] Thanks to the collaboration of many different scholars working on Greek and Latin inscriptions, *EpiDoc* (Epigraphic Documents; *http://sourceforge.net/p/epidoc/wiki/Home/)* has been established as a robust system for what regards the representation and the encoding of epigraphic or papyrological texts in digital form (cf. Bodard, 2010). *EpiDoc* adopts a subset of the XML defined by the TEI standard for the digital representation of texts, which is now widely used in the humanities. This flexible system allows not only the transcribing of a Greek or Latin text, but also, for instance, the encoding of its translation, description, and other pieces of information such as dating, history of the inscription, bibliography, and the object on which the text is written.

Latin by combining the investigation of diastratic (Clackson, 2011a; Adams, 2013), diatopic (Herman, 1990; Adams, 2007), diaphasic (Kruschwitz and Halla-aho, 2007; Kruschwitz, 2015; Ferri and Probert, 2010), and diamesic variation.

However, none of the corpora illustrated above allows researchers to directly access specific information about relevant linguistic variation phenomena, and they do not satisfactorily meet the needs of the linguist to study Latin epigraphic texts from a variationist perspective. In particular, as already stated above (§ 2.1), in all of them queries can be performed using a token-level keyword search by entering single words or set of words or letters, and this requires prior knowledge of what to search for. In addition, one cannot always be sure that the digital editions of the texts are free from emendations and standardizations of those aberrant forms, misspellings, and spelling variants that are of primary relevance for the linguist.

Thus, in order to systematically address the massive (ortho)graphic and linguistic variation observable in Latin inscriptions, differently designed tools are necessary. This is the reason why *CLaSSES*, while providing annotation for both extra- and meta-linguistic data (§ 4.1), also provides fine-grained linguistic information about specific spelling variants that can be regarded as clues for phonetic-phonological and morpho-phonological variation (cf. § 4.2). Another database that is designed to be a helpful tool in the study of linguistic (diatopic) variation is the Computerized Historical Linguistic Database of the Latin Inscriptions of the Imperial Age (*LLDB*)[11], a comprehensive digital resource for the Vulgar Latin inscriptions from the Roman provinces (Adamik, 2012). More than 87,800 spellings that diverge from the Classical norm are collected in *LLDB* and they are accurately classified according to a wide range of phonetico-phonological, morphological, and syntactic phenomena. Moreover, each form is richly annotated with extra- and meta-linguistic information including findspot, dating, type of inscription (e.g. Christian or non-Christian, prose or verse, private or official), type of object, comments on issues concerning the reading of the texts (e.g. presence of fractures on the object, etc.), and relevant bibliography. The search interface makes it possible to perform simple and advanced queries by combining an unlimited number of search criteria and by using Boolean operators. However, it has to be noted that this resource has been

[11] Cf. *http://lldb.elte.hu/* [accessed on 20.02.2020]. The database is a revised and upgraded version of József Herman's Late Latin Data Base, hence the acronym.

designed in order to meet the requirements of Herman's (2000, for the last version) approach to the investigation of language variation. According to him, divergent spellings can be assumed as representative of diatopic variation only if their relative frequency is expressed as a percentage against the total number of other linguistically relevant divergent spellings (for an updated discussion of the methodological issues, see Tamponi, 2020: 24-26). As a consequence, in *LLDB* it is possible to elicit lists of misspelled forms, but they cannot be checked against the total amount of the corresponding Classical spellings.

2.4. *Other non-literary texts; papyri, wooden tablets, and Medieval charters*

A few last words are due for the digital editions of other non-literary texts (such as papyri and letters of correspondence) that can be a valuable source for variationist analysis. Papyri.info, is an extensive digital text collection of Greek and Latin documentary papyri dating from the 4th century BC to the 8th century AD, in large part from Egypt. The resource is based on the Papyrological Navigator (*PN*), a tool that aggregates three major databases of documentary papyri: the Duke Databank of Documentary Papyri (*DDbDP*), the Heidelberger Gesamtverzeichnis der griechischen Papyruskunden Ägyptens (*HGV*), and the University of Michigan Advanced Papyrological Information System (*APIS*)[12]. The main bibliographical database for papyrological research, the Bibliographie Papyrologique (*BP*), is also integrated. The texts, coming from the *DDbDP*, have been converted in *EpiDoc* and are now integrated and merged with metadata and images drawn from the *HGV* and the *APIS* databases. The archive currently includes 56,779 texts (in addition, there are 29,867 records with metadata only). The Navigator allows both simple and complex string-searching and the search can be refined by adding further criteria (series and collection, provenance, dating, language, etc.). Annotations of linguistic phenomena are lacking, so that the texts cannot be queried in this way, but it is worth mentioning that a corpus of Greek texts exported from papyri.info has been enriched with linguistic information as part of the SEMATIA Project (Linguistic Annotation of the Greek

[12] Cf. *http://papyri.info/* [accessed on 20.02.2020]. Another important tool that is available in papyri.info, is the Papyrological Editor (*PE*), which enables users to contribute to the collection by entering new texts and metadata, or editing those already existing.

Documentary Papyri - Detecting and Determining Contact-Induced, Dialectal and Stylistic Variation) of the University of Helsinki[13]. The result is an extensively annotated corpus that enables the comparison between the misspellings and spelling variants of the scribes' original text and the standard Greek, as well as the analysis of the morpho-syntactic structures of the texts. For a corpus of linguistically annotated Latin papyri, see below (§ 3.3) the section of *CLaSSeS Egypt and Eastern Mediterranean*.

Another important collection of Latin first-hand texts is the digital publication of the ink-written wooden tablets from the Roman garrison of Vindolanda, dating between the 1st and the 3rd century AD. The documents include private correspondence, military reports, accounts, and other informal or non-literary writings. The online edition is hosted by two separate websites: *http://vindolanda.csad.ox.ac.uk* for the tablets published in Bowman and Thomas (1983) and Bowman and Thomas (1994), *http://vto2.classics.ox.ac.uk* for the tablets published in Bowman and Thomas (2003) and the earlier publications. Simple word queries can be performed by means of the 'Latin text search', while other information (subjects, categories and types of documents, people, places, military terms, archaeological context) can be accessed through the 'General text search' facility or through browsing. Every text is transcribed, translated, provided with a photo and an accurate description with particular focus on the palaeographic aspects. Specific linguistic annotation is missing also in this case, but for its implementation for 762 ink-written tablets as part of *CLaSSeS*, see § 3.2 *Roman Britain* below.

While none of the corpora illustrated in this section is specifically designed for linguistic analysis, a notable exception is the Late Latin Charter Treebanks (*LLCT*), which is developed for the research of the non-literary Latin of the Early Middle Ages (Korkiakangas, 2020, and references therein). The *LLCT* treebank is a set of three morphologically and syntactically annotated corpora (*LLCT1*, *LLCT2*, *LLCT3*), which also feature a textual annotation layer that indicates abbreviated and restored words. *LLCT1* and *LLCT2* are now completely accessible online[14]: the former includes 225,834 tokens distributed within 519 charters written in Tuscany between 714 and 869 AD; the latter includes 257,819 tokens in 521 Tuscan charters between 774 and 897 AD. *LLCT3*, under construction, is going to contain

[13] Cf. *https://sematia.hum.helsinki.fi*.

[14] Cf. *https://zenodo.org/record/3633607#.XjU4lSNS9EY* (for *LLCT1*) and *https://zenodo.org/record/3633614#.XjU6zCN7lEY* (for *LLCT2*).

ca. 110,400 tokens in 221 charters written in Tuscany as well as in several locations in northern and southern Italy between 721 and 1000 AD. As the lemmatization and grammatical parsing of traditional treebanks[15] is mainly based on texts of Classical authors (for an overview of Latin lemmatizers and morphological analyzers, see Celano, 2020), in *LLCT* particular attention is paid to the lemmatization and additional annotation of all those non-classical and late forms that are typical of non-literary Early Medieval Latin.

3. *Materials*

CLaSSES is structured in four different sections, whose contents are hereafter described with reference to the kind of material, dating, and area of provenance: *Rome and Italy* (§ 3.1), *Roman Britain* (§ 3.2), *Egypt and Eastern Mediterranean* (§ 3.3), and *Sardinia* (§ 3.4). These sections can be also accessed from an interactive map, which shows the number and the geographic distribution of the inscriptions included in the database. The criteria of tokenization, lemmatization, as well as those of linguistic, meta-linguistic and extra-linguistic annotation are illustrated in § 4 below.

3.1. *Rome and Italy*

The first section, *Rome and Italy*, is a collection of 1,250 Latin inscriptions (for a total number of 11,804 tokens), dating between the 6th century BC and the 1st century AD, mainly from Rome and Central Italy. The inscriptions belong to five different textual typologies (*tituli honorarii*, *tituli sepulcrales*, *instrumenta domestica*, *tituli sacri publici*, and *tituli sacri privati*; cf. § 4.1 for the criteria of classification), and their texts have been retrieved from the following editions: Lommatzsch (1918, *Hrsg.*; 1931, *Hrsg.*; 1943, *Hrsg.*), Degrassi and Krummrey (1986, eds.), Dressel (1899 [1969]), Gordon and Gordon (1958), Panciera *et al.* (1991), Degrassi (1957-1963), Wachter (1987), and Warmington (1940)[16].

[15] Cf. the Latin Dependency Treebanks (*LDT*, *https://perseusdl.github.io/treebank_data/*), the *PROIEL* treebanks (*https://proiel.github.io*), and the Index Thomisticus Treebank (*IT-TB*, *https://itreebank.marginalia.it*).

[16] Note that, among the available material, not every inscription is significant for linguistic studies. As a consequence, the following texts have been excluded: (i) legal texts, since they are generally prone to archaisms; (ii) too short (single letters, initials) or fragmentary inscriptions; (iii) inscriptions from the necropolis of Praeneste, as they contain only anthroponyms in the nominative form.

The study of spelling variants in archaic and early epigraphy is of particular relevance for the investigation of the long-lasting process of formal codification of the language that led to what is currently labelled 'Classical Latin'. The contrastive analysis between the language of these inscriptions and that which became an established standard with fixed rules and forms, is representative of the fundamental process of selection, regularization, and reduction of variation underlying the ideology of *Latinitas* "correct Latin", which was progressively elaborated by grammarians, rhetoricians, poets, and prose writers between the final decades of the Republic and the early Empire (Poli, 1999; Clackson and Horrocks, 2007: 130-182; Clackson, 2011a; 2011b; Cuzzolin and Haverling, 2009; Mancini, 2005; 2006).

In the absence of an established standard and as a consequence of specific and particular issues of single inscriptions, the texts of this period may raise problems with their reading and with the linguistic interpretation of their forms. In such cases, the numerous readings that have been proposed so far by scholars have been compared in order to guarantee the most reliable and updated philological accuracy.

3.2. *Roman Britain*

The section *Roman Britain* has, so far, an assemblage of 762 ink-written tablets (for a total number of 11,446 tokens) from the auxiliary fort Vindolanda just south of Hadrian's Wall, dating between the 1st and the 3rd century AD. The inscriptions belong to ten different textual typologies: military reports, *commeatus*, *numera*, *memorandum*, *commendatio*, male / female correspondence, *literaria*, miscellany, and *descripta* (cf. § 4.1 for the criteria of classification). For this section, the inscriptions were collected from the following corpora and online resources: Bowman and Thomas (1983; 1994; 2003), Bowman, Thomas and Tomlin (2010), Bowman, Thomas and Tomlin (2011), *http://vindolanda.csad.ox.ac.uk/*, *http://vto2.classics.ox.ac.uk/* (cf. above, § 2.4).

Since Adams (1995) the language of the Vindolanda writing-tablets has attracted the attention of scholars working on language variation and contact. On the one hand it is possible to identify different degrees of literacy between the texts written by the prefects and their scribes, and those written by other people with poorer competence, whose misspellings allow linguistic considerations (Cotugno, 2015; Cotugno and Marotta,

2017). On the other hand, in this military post Latin was used by auxiliary troops coming mainly from Gallia Belgica, i.e. Celto-Germanic people whose Latin writings may bear tell-tale signs of second-language learning (Cotugno, 2018).

3.3. *Egypt and Eastern Mediterranean*

This section has a collection of 220 documentary letters (for a total number of 9,224 tokens) written on papyri and ostraka from Africa Proconsularis, Aegyptus, Palestine, and Syria, dating between the 1st and the 6th century AD. Two different textual typologies have been distinguished on the basis of the epistolary genre: formal (i.e. public) and informal (i.e. private) letters. The documents from these areas were retrieved from the following editions: Cugusi (1992a; 1992b; 2002) and Marichal (1992).

Greek remained the lingua franca of all the eastern regions of the Empire and it was used as such also by the Romans, and the Latin-speaking population in these areas largely consisted of not locally born Latin speakers, but «mobile personnel, who would no doubt adopt 'regional' usages as they came and went» (Adams, 2003: 525). As a consequence, this corpus of documentary letters, which was the work of a variety of bilingual (and possibly bi-literate) scribes[17], is of particular interest both for the study of regional variation and for the study of linguistic and graphemic interference between Latin and Greek (Barchi, 2019).

3.4. *Sardinia*

The last section contains 1,184 inscriptions (for a total number of 14,413 tokens) from Sardinia, dating between the 1st century BC and the beginning of the 7th century AD. In line with the criteria adopted for the section *Rome and Italy*, the following textual typologies have been identified: *tituli honorarii*, *tituli sepulcrales*, *tituli sacri publici*, *tituli sacri privati*, *instrumenta domestica*; the supplementary category *military diplomas* has been added. The reference editions for the texts are Mommsen (1883, *Hrsg.*), Ihm (1899), Sotgiu (1961; 1968; 1988), Corda (1999), Floris (2005).

[17] Cf. the well-known case of Claudius Terentianus, illustrated (among others) in Adams (2003: 527-637, 741-750 and *passim*).

As Roman Sardinia was a multi-faceted community of speakers, a quantitative analysis of the surviving Latin inscriptions can provide insights into the dynamics of diatopic variation and interference. In particular, it is likely it will sustain quantitative evidence backing the traditional hypothesis that acknowledges a number of common features between African Latin and the Latin of Sardinia (Fanciullo, 1992; Lupinu, 2003; Lorenzetti and Schirru, 2010; Loporcaro, 2015: 48 ff.), as well as casting some light on the specific evolution of the Sardo-Romance varieties among the Romance languages (Tamponi, 2020).

4. *Corpus annotation*

As just described, *CLaSSES* includes 3,416 Latin documents in digitized form. As a preliminary operation for the creation of the database, all texts have been automatically tokenized, i.e. broken into a sequence of words and units of punctuation (for a total number of 46,887 tokens).

Each token of the corpus is univocally associated with a token-ID, i.e. a short string of alphanumeric characters that provide basic information: the source of the text, the number of the inscription, and the position in which the token occurs within the inscription (e.g. BTT-118-1 refers to Bowman, Thomas and Tomlin's edition of the Vindolanda writing-tablets, publication number 118, and first word of the text).

After tokenization, all words of the corpus (also abbreviated and incomplete forms that could be fully understood) have been lemmatized. This operation was conducted manually, due to the high frequency in letters and inscriptions of abbreviated, incomplete, and misspelled words that could not be easily processed by automatic tools.

Once tokenized and lemmatized, a rich linguistic, meta-linguistic, and extra-linguistic annotation has been added to the texts, as described in the following paragraphs (cf. also De Felice *et al.*, 2015). Data were recorded in a tabular form in Excel worksheets by four expert annotators (cf. Section *Acknowledgments*), who worked separately on the different subsections of *CLaSSES*. All the data collected were carefully cross-checked by other annotators and researchers involved in the project (disagreements were collaboratively discussed to reach consensus), before being converted into a database that can now be freely accessed from the *CLaSSES* website (cf. § 5).

4.1. *Extra-linguistic and meta-linguistic annotation*

Place of provenance and dating. Extra-linguistic information related to the place of provenance and dating of each document included in the database has been annotated (these data were derived from the sources from which the texts were retrieved). Places of provenance can be grouped into four main areas: Rome and peninsular Italy, Sardinia, Egypt and Eastern Mediterranean, and Roman Britain. The dating of the collected documents spans from the 6th-5th century BC of some inscriptions from Central Italy to the 5th-7th century AD of some Egyptian papyri and Sardinian texts (cf. § 3).

Text type. Each text has also been classified according to its typology. Among the epigraphic texts collected in the sections *Rome and Italy* and *Sardinia* we find *tituli honorarii* (honorary inscriptions dedicated by public figures and monumental inscriptions), *tituli sepulcrales* (commemorative inscriptions and epitaphs), *instrumenta domestica* (inscriptions on everyday objects), *tituli sacri publici* (votive inscriptions dedicated by public figures), *tituli sacri privati* (votive inscriptions dedicated by private customers), and *military diplomas* (this last category is used only for Sardinian texts, to classify personal legal documents on bronze tablets that contain a copy of imperial constitutions by which Roman citizenship and *conubium* were granted to veterans of the auxiliary army units, the fleet and the Praetorian Guard).

Vindolanda's tablets (section *Roman Britain*) may be classified as *military reports* (communications between officers regarding the activity of the garrison), *commeatus* (applications of leave to the prefect of the cohort), *memoranda* (short communications left by one garrison to the other), *commendationes* (letters of recommendations), *numera* (accounts of various types), *literaria* (writing exercises), *male/female correspondence*, *miscellany* (tablets of uncertain attribution), and *descripta* (tablets with a very faded text, for which there are doubts about their reconstruction).

Finally, the letters collected in the section *Egypt and Eastern Mediterranean* have been classified as either *formal* (i.e. public) or *informal* (i.e. private letters of information).

Most of the categories adopted for classifying the text types were derived from the original sources of the digitized texts, but, in many cases, annotators created specific labels to provide a more fine-grained classification (for instance, making a distinction, within the group of the inscriptions tra-

ditionally classified in the *CIL* as *tituli sacri*, between *tituli sacri privati* and *tituli sacri publici*; cf. also Donati, 2015).

Graphic form. The epigraphic texts, tablets, and letters collected in *CLaSSES* rarely consist of well-written and fully readable words; rather, they often present faint or missing letters betrayed by the conservation status of the support, or incomplete forms (initials, abbreviations). Therefore, each token of the corpus has been also classified according to its graphic form. For this level of annotation, we distinguish the following categories: *complete words*; *abbreviations*, for every kind of shortening (e.g. BTT-135-5 COH for COHORS), including personal name initials; *incomplete words*, for words partly integrated by editors (e.g. ILSARD-I-388-37 AURE[LIO]) or impossible to integrate (e.g. CEL-I-5-2 GLAU[); *words completely integrated by editors* (e.g. BTT-257-2 [CERIALI]); *presumed misspellings* (e.g. CEL-I-1-416 SITULUS for TITULUS); *uncertain words*, for words that cannot be interpreted, not even in their graphical form (e.g. CIL-I^2-59-9 STRIANDO); *numbers*; *symbols*, only in the sections *Roman Britain* and *Sardinia*, for non-alphabetical signs (that are presented in the database not as graphic signs, but with an indication of their meaning between brackets, e.g. BTT-138-6 and EE-VIII-710-11 SYMBOL(CENTURIAE)); *lacunae*, i.e. gaps in the inscription (*lacunae* are identified by the string [...] and they are considered to be tokens, since they occupy a specific position within the texts, and they actually exist in their critical editions).

Language. Even if the documents which compose the corpus *CLaSSES* are primarily written in Latin, they sometimes include foreign words. Therefore, we distinguished Latin forms from words belonging to other languages, manually annotated as *Greek*, *Oscan*, *Umbrian*, *Etruscan*, *Iberian*, *Neo-Punic*, *Semitic*, *Coptic*, *Hebrew*, *Egyptian*, and *Persian*. Moreover, mixed forms are marked as *hybrid* (e.g. CIL-I^2-553-2 ALIXENTROM, a Greek loanword in a Latin form with Etruscan phono-morphological interferences), whereas those of unknown language are marked as *unknown* (e.g. CEL-I-150-39 ATESTAS).

Author/addressee. Only for the section *Roman Britain*, containing letters from Vindolanda, did we choose to also annotate the author of the texts and his/her addressees when the identity of these persons is known. For instance, the tablet BTT-233 is written by *Cerialis* and addressed to *Aelius Brocchus*.

4.2. *Linguistic annotation of non-classical variants*

The most relevant part of the annotation process, which provides the corpus with a rich set of qualitative data, is the result of an accurate and in-depth linguistic analysis of the collected documents. The purpose of this annotation is (i) to identify non-classical variants, i.e. all words that deviate from Classical Latin from a purely (ortho-)graphic point of view (as described in § 1; see also Marotta, 2015; 2016), and (ii) to classify non-classical variants according to the kind of variation phenomenon involved. Therefore, first annotators manually identified all words that clearly do not belong to the classical literary language (e.g. DEDE instead of classical DEDIT; MENERVAI instead of classical MINERVAE) and marked them as *non-classical* (tot. 3,838, i.e. 8,2% of tokens in the four sub-sections of *CLaSSES*). Then, they associated each non-classical form with its corresponding classical form (e.g. nom. sg. CORNELIO, non-classical - CORNELIUS, classical). Finally, all non-classical variants were classified according to the type of variation phenomena that distinguish them from the corresponding classical equivalents. More precisely, such variation phenomena may regard the vowel or consonant system, as well as morpho-phonology (when variation occurs in morphological endings of words). The most relevant phenomena annotated for vowels are the following:

- vowel alternations (CIL-I²-2909-4 MENERVA for MINERVA; BTT-206-34 SENICIO for SENECIO);
- phenomena related to the notation of vowel length, such as vowel doubling (CIL-I²-365-11 VOOTUM for VOTUM), *apex* (CEL-I-8-33 SUÓ for SUO), and *I longa* (BTT-297-9 FECI for FECI);
- omission of vowel (CIL-I²-37-10 VICESMA for VICESIMA; CIL-X-7756-28 OCLOS for OCULOS) and insertion of vowel (BTT-187-15 CRISPIA for CRISPA);
- phenomena related to diphthongs (such as <E> for Classical <AE> in CEL-I-157-17 ETATIS).

The main phenomena related to consonants can be summarized as follows:

- omission of final consonant (CIL-I²-8-2 CORNELIO for CORNELIUS; CIL-X-7809-15 ANNU for ANNUM);
- omission of nasal before consonants (CEL-I-177-8 PRAESES for PRAESENS; BTT-609-39 SACTIUS for SANCTIUS);

– assimilation (CEL-I-77-47 MASSIPIUM for MARSIPIUM);
– double *pro* single consonant (CIL-I²-16-1 [P]AULLA for [P]AULA) and single *pro* double consonant (CEL-I-234-37 QUATUOR for QUATTUOR);
– <B>/<V> confusion (CIL-X-7990-16 BIXIT for VIXIT; CIL-X-7619-11 VENE for BENE).

Most of the categories just presented are further articulated into sub-categories, in order to allow a more fine-grained classification of variation phenomena; for instance, for vowel alternations we annotated as two separate phenomena (i) <I>, /ī/ = <E> and (ii) <I>, /ī/ = <E>.

If non-classical variants occur in morpho-phonological position (generally, in word endings), we also annotated the special ending attested, such as the *-e* ending of the dative singular of the first declension (CEL-I-146-57 MEE for MEAE), the *-os* and *-o* endings of the nominative singular of the second declension (CIL-I²-406b-2 CANOLEIOS and CIL-I²-408-2 CANOLEIO for CANOLEIUS), the *-om* ending of the accusative singular of the second declension (CIL-I²-403-8 LOCOM for LOCUM), or the *-et* ending of the 3rd person of the perfect (CIL-I²-365-12 DEDET for DEDIT; CIL-X-7632-12 FECET for FECIT).

5. *Search interface*

The open-access search interface currently available on the *CLaSSES* website (*http://classes-latin-linguistics.fileli.unipi.it*) has been specifically developed to explore the corpus, to perform queries on it, and to access the fine-grained linguistic annotation conducted on texts.

Basic queries can be made by clicking the *Search* button from the top menu of the website and by selecting the sub-corpus of interest: *Rome and Italy*, *Sardinia*, *Roman Britain*, or *Egypt and Eastern Mediterranean* (documents can be also selected from the map in the *Homepage*, which shows the geographic distribution and the number of the texts included in the database). It is also possible to query the whole corpus, by selecting *Cross-corpora*. Once the section of interest is selected and the search interface accessed, the entire (sub-)corpus is displayed in a vertical column, with one token per row. Most data annotated for each token are reported in multiple columns in a tabular format: its ID (containing information about the publication number of the inscription or letter and the source from which the text is derived);

its lemma, language, and graphic form; its classification as either a classical word or a non-classical variant; the typology of the inscription or the letter which the token belongs to, its place of provenance, its dating; the author and addressee of the letters (only for the section *Roman Britain*); and the support material of the document (only for the section *Egypt and Eastern Mediterranean*).

It is possible to perform simple queries on the corpus, either by searching for a specific form (the use of 'wildcard' characters is supported), or by using and combining the filters at the top of each column (for instance, to visualize only classical or non-classical forms, to filter results per publication number, lemma, language, graphic form, and place of provenance, etc.). With the *Advanced search* functionality, users can select more than one option for a search filter (e.g. for language: Latin AND Greek AND Hybrid); most importantly, it is also possible to search for specific linguistic phenomena annotated for vowels, consonants, or morphophonology. Finally, the export options in the *Search* page allow exporting the data in different formats (CSV, Text, Excel 1995+, Excel 2007+), at any moment.

The two columns on the rightmost part of the *Search* page (*Text* and *More Info*) allow access to further information. By clicking on the two symbols present in the *Text* column, it is possible to visualize the immediate linguistic context of each form of the corpus (5 words before/after) and to read the entire text of the document. By clicking on the symbol present in the *More Info* column, a new page will open containing all data annotated for a given form: token ID, language, graphic form, lemma, classical/non-classical classification, text typology, place of provenance, dating, author, addressee, support material, linguistic context, and entire inscription; in case of non-classical form, the equivalent classical form is reported (for instance, CONSUL for non-classical COSOL). At the end of this page, the variation linguistic phenomena individuated for non-classical forms are reported.

6. *Conclusions*

In conclusion, *CLaSSES* aims at being an additional digital resource for academic scholarship which is interested in carrying out variationist studies on the non-literary documentation of the Latin language.

Of course, the database is built on a reference corpus of texts which is not on a par either with other available, extensive digital epigraphic collections, or with the existing treebanks and lemmatizers that are based on large repertoires of literary texts (many of which are described in the other papers of this special issue of *Studi e Saggi Linguistici*). A full coverage of the non-literary documents is not its purpose, after all.

Rather, the corpus is designed for the investigation of orthographic variants in non-literary Latin texts of various ages and provenance. Due to their nature, these sources allow us to draw relevant data on the phonological and morpho-phonological domains, which other available digital tools do not provide with such fine-grained annotation.

CLaSSES relies on single and coherent corpora of texts, in which the annotation of orthographic variation is systematically cross-referenced with the meta-linguistic information. Such a correlation between linguistic data and extra-linguistic variables can provide reliable clues in order to perform diachronic, diatopic, and diaphasic analyses, which may hopefully cast some further light on the sociolinguistic variation within the Latin language.

Acknowledgments

CLaSSES is the product of many people's research. In particular, the section *Rome and Italy* was developed by Lucia Tamponi and Margherita Donati; the section *Roman Britain* by Francesca Cotugno; the section *Egypt and Eastern Mediterranean* by Serena Barchi; the section *Sardinia* by Lucia Tamponi.

We would like to thank these young researchers for their invaluable work: without them our project would never have seen the light. The Authors of the present article have supervised every entry in the database. Irene De Felice implemented the tokenization and contributed to the annotation of all the materials available in *CLaSSES*. Thanks are also due to Stefano Dei Rossi for his constant technical assistance.

The valuable comments and observations made by the two anonymous referees helped us to improve the overall design of the paper. We are very grateful to them for their appropriate and constructive suggestions.

The present paper was conceived and discussed by the four authors in agreement. For academic reasons only, the scientific responsibility is attributed as follows: §§ 1 and 6 to Giovanna Marotta; § 2 to Francesco Rovai; § 3 to Lucia Tamponi; §§ 4 and 5 to Irene De Felice.

References

Adamik, B. (2012), *In search of the regional diversification of Latin: Some methodological considerations in employing the inscriptional evidence*, in Biville, F., Lhomme, M. and Vallat, D. (2012, *éds.*), *Latin vulgaire - Latin tardif IX. Actes du IX^e^ colloque international sur le latin vulgaire et tardif, Lyon, 6-9 septembre 2009*, Maison de l'Orient et de la Méditerranée 'Jean Pouilloux', Lyon, pp. 123-139.

Adams, J.N. (1995), *The Language of the Vindolanda Writing Tablets: An Interim Report*, Cambridge University Press, Cambridge.

Adams, J.N. (2003), *Bilingualism and the Latin Language*, Cambridge University Press, Cambridge.

Adams, J.N. (2007), *The Regional Diversification of Latin 200 BC-AD 600*, Cambridge University Press, Cambridge.

Adams, J.N. (2013), *Social Variation and the Latin Language*, Cambridge University Press, Cambridge.

Amato, G., Bollettieri, P., Gennaro, C., Manghi, P., Mannocci, A., Zoppi, F., Casarosa, V. and Falchi, F. (2013), *AIM infrastructure specification* [available online at *https://www.eagle-network.eu/wp-content/uploads/2013/06/EAGLE_D4.1_AIM_Infrastructure_Specification_update.pdf*].

Barchi, S. (2019), *On vowel prosthesis before sC in Substandard Latin and Koine Greek: a synoptic review*, in «Studi e Saggi Linguistici», 57, 2, pp. 45-82.

Biville, F., Decourt, J.C. and Rougemont, G. (2008, *éds.*), *Bilinguisme gréco-latin et épigraphie,* Maison de l'Orient et de la Méditerranée 'J. Pouilloux', Lyon.

Bodard, G. (2010), *EpiDoc: Epigraphic Documents in XML for publication and interchange*, in Feraudi-Gruénais, F. (2010, ed.), *Latin on Stone: Epigraphic Research and Electronic Archives*, Lexington Books, Lanham, pp. 101-118.

Bowman, A.K. and Thomas, J.D. (1983), *Vindolanda: The Latin Writing Tablets*, Society for the Promotion of Roman Studies, London.

Bowman, A.K. and Thomas, J.D. (1994), *The Vindolanda Writing Tablets (*Tabulae Vindolandenses *II)*, British Museum Press, London.

Bowman, A.K. and Thomas, J.D. (2003), *The Vindolanda Writing Tablets (*Tabulae Vindolandenses *III)*, British Museum Press, London.

Bowman, A.K., Thomas, J.D. and Tomlin, R.S.O. (2010), *The Vindolanda writing tablets (*Tabulae Vindolandenses *IV part 1)*, in «Britannia», 41, pp. 187-224.

Bowman, A.K., Thomas, J.D. and Tomlin, R.S.O. (2011), *The Vindolanda writing tablets (*Tabulae Vindolandenses *IV part 2)*, in «Britannia», 42, pp. 113-144.

Caldelli, M.L., Orlandi, S., Blandino, V., Chiaraluce, V., Pulcinelli, L. and Vella, A. (2014), EDR - *Effetti collaterali*, in «Scienze dell'Antichità», 20, 1, pp. 267-289.

Campanile, E. (1971), *Due studi sul latino volgare*, in «L'Italia Dialettale», 34, pp. 1-64.

Celano, G. (2020), *Lemmatization and morphological analysis for the Latin Dependency Treebank*, in «Studi e Saggi Linguistici», 58, 1, pp. 21-38.

Clackson, J. (2011a), *Latin inscriptions and documents*, in Clackson, J. (2011, ed.), *A Companion to the Latin Language*, Wiley / Blackwell, Chichester / Malden, pp. 29-39.

Clackson, J. (2011b), *The social dialects of Latin*, in Clackson, J. (2011, ed.), *A Companion to the Latin Language*, Wiley / Blackwell, Chichester / Malden, pp. 505-526.

Clackson, J. and Horrocks, G. (2007), *The Blackwell History of the Latin Language*, Blackwell, Malden.

Conde-Silvestre, N. and Hernández-Campoy, J.M. (2012), *Introduction*, in Hernández-Campoy, J.M. and Conde-Silvestre, N. (2012, eds.), *The Handbook of Historical Sociolinguistics*, Oxford University Press, Oxford, pp. 1-7.

Consani, C. (2016), *Fenomeni di contatto a livello di discorso e di sistema nella Cipro ellenistica (Kafizin) e le tendenze di "lunga durata"*, in «Linguarum Varietas», 5, pp. 51-65.

Corda, A.M. (1999), *Le iscrizioni cristiane della Sardegna anteriori al VII secolo*, Pontificio Istituto di Archeologia Cristiana, Città del Vaticano.

Cordell, R. (2015), *Reprinting, circulation, and the network author in antebellum newspapers*, in «American Literary History», 27, 3, pp. 417-445.

Cotugno, F. (2015), I longa *in iato nel* Corpus Vindolandense, in «Studi e Saggi Linguistici», 53, 2, pp. 189-206.

Cotugno, F. (2018), *The Linguistic Variation of Latin in Roman Britain*, PhD thesis, University of Pisa.

Cotugno, F. and Marotta, G. (2017), *Geminated consonants in the Vindolanda tablets. Empirical data and sociolinguistics remarks*, in Molinelli, P. (2017, ed.), *Language and Identity in Multilingual Mediterranean Settings. Challenges for Historical Sociolinguistics*, Mouton de Gruyter, Berlin, pp. 269-288.

Cugusi, P. (1992a), *Corpus Epistularum Latinarum Papyris Ostracis Tabulis servatarum*. Vol. 1: *Textus*, Gonnelli, Firenze.

Cugusi, P. (1992b), *Corpus Epistularum Latinarum Papyris Ostracis Tabulis servatarum*. Vol. 2: *Commentarius*, Gonnelli, Firenze.

Cugusi, P. (2002), *Corpus Epistularum Latinarum Papyris Ostracis Tabulis servatarum*. Vol. 3: *Addenda, Corrigenda, Indices rerum, Index verborum omnium*, Gonnelli, Firenze.

Cuzzolin, P. and Haverling, G. (2009), *Syntax, sociolinguistics, and literary genres*, in Baldi, P. and Cuzzolin, P. (2009, eds.), *New Perspectives on Historical Latin Syntax: Syntax of the Sentence*, De Gruyter, Berlin / New York, pp. 19-64.

De Felice, I., Donati, M. and Marotta, G. (2015), *CLaSSES: A new digital resource for Latin epigraphy*, in «Italian Journal of Computational Linguistics», 1, 1, pp. 119-130.

Degrassi, A. (1957-1963), *Inscriptiones latinae liberae rei publicae*. 2 Voll., La Nuova Italia, Firenze.

Degrassi, A. and Krummrey, J. (1986, *Hrsg.*), *CIL I² 2,4: Addenda tertia. Addenda ad inscriptiones vetustissimas*, De Gruyter, Berlin.

Dickey, E. and Chahoud, A. (2010, eds.), *Colloquial and Literary Latin*, Cambridge University Press, Cambridge.

Donati, M. (2015), *Variazione e tipologia testuale nel corpus epigrafico* CLaSSES I, in «Studi e Saggi Linguistici», 53, 2, pp. 21-38.

Dressel, H. (1899 [1969]), *CIL XV, 2: Inscriptiones urbis Romae Latinae. Instrumentum domesticum*, Reimer, Berlin.

Eck, W. and Funke, P. (2014, *Hrsg.*), *Öffentlichkeit - Monument - Text. XIV Congressus Internationalis Epigraphiae Graecae et Latinae Akten*, De Gruyter, Berlin.

Elliott, T. (2015), *Epigraphy and digital resources*, in Bruun, C. and Edmondson, J. (2015, eds.), *The Oxford Handbook of Roman Epigraphy*, Oxford University Press, Oxford / New York, pp. 78-85.

Fanciullo, F. (1992), *Un capitolo della Romania submersa: il latino africano*, in Kremer, D. (1992, *éd.*), *Actes du XVIII^e Congrès International de Linguistique et de Philologie Romanes (Trier 1986)*. Vol. 1, Niemeyer, Tubingen, pp. 162-187.

Feraudi-Gruénais, F. (2010), *An inventory of the main archives of Latin inscriptions*, in Feraudi-Gruénais, F. (2010, ed.), *Latin on Stone: Epigraphic Research and Electronic Archives*, Lexington Books, Lanham, pp. 157-160.

Ferri, R. and Probert, P. (2010), *Roman authors on colloquial language*, in Dickey, E. and Chahoud, A. (2010, eds.), *Colloquial and Literary Latin*, Cambridge University Press, Cambridge, pp. 12-41.

Floris, P. (2005), *Le iscrizioni funerarie pagane di Karales*, Edizioni AV, Cagliari.

Gordon, J. and Gordon, A. (1958), *Album of Dated Latin Inscriptions*. Vol. 1, University of California Press, Berkeley / Los Angeles.

Herman, J. (1985), *Témoignage des inscriptions latines et préhistoire des langues romanes: le cas de la Sardaigne*, in Deanović, M. (1985, *éd.*), *Mélanges de linguistique dédiés à la mémoire de Petar Skok (1881-1956)*, Jugoslavenska Akademija Znanosti i Umjetnosti, Zagreb, pp. 207-216.

Herman, J. (1990), *Du latin aux langues romanes. Études de linguistique historique*, Niemeyer, Tübingen.

Herman, J. (2000), *Differenze territoriali nel latino parlato dell'Italia: un contributo preliminare*, in Herman, J. and Marinetti, A. (2000, *a cura di*), *La preistoria dell'italiano. Atti della Tavola Rotonda di Linguistica Storica. Università Ca' Foscari di Venezia, 11-13 giugno 1998*, Niemeyer, Tübingen, pp. 123-135.

Ihm, M. (1899), *Additamenta ad Corporis vol. IX et X*, in *Ephemeris epigraphica. Corporis inscriptionum Latinarum supplementum, edita iussu Instituti archaeologici Romani*. Vol. 8, Reimer, Berlin, pp. 1-22.

Korkiakangas, T. (2020), *Theoretical and pragmatic considerations on the lemmatization of non-standard Early Medieval Latin charters*, in «Studi e Saggi Linguistici», 58, 1, pp. 67-94.

Kruschwitz, P. and Halla-aho, H. (2007), *The Pompeian wall inscriptions and the Latin language: a critical reappraisal*, in «Arctos», 41, pp. 31-49.

Kruschwitz, P. (2015), *Linguistic variation, language change, and Latin inscriptions,* in Bruun, C. and Edmondson, J. (2015, eds.), *The Oxford Handbook of Roman Epigraphy*, Oxford University Press, Oxford / New York, pp. 721-743.

Lommatzsch, E. (1918, *Hrsg.*), *CIL I² 2,1: Inscriptiones vetustissimae*, De Gruyter, Berlin.

Lommatzsch, E. (1931, *Hrsg.*), *CIL I² 2,2: Addenda Nummi Indices. Addenda ad inscriptiones vetustissimas*, De Gruyter, Berlin.

Lommatzsch, E. (1943, *Hrsg.*), *CIL I² 2,3: Addenda altera Indices. Addenda ad inscriptiones vetustissimas*, De Gruyter, Berlin.

Loporcaro, M. (2015), *Vowel Length from Latin to Romance*, Oxford University Press, Oxford.

Lorenzetti, L. and Schirru, G. (2010), *Un indizio della conservazione di /k/ dinanzi a vocale anteriore nell'epigrafia cristiana di Tripolitania*, in Tantillo, I. and Bigi, F. (2010, *a cura di*), *Leptis Magna. Una città e le sue iscrizioni in epoca tardoromana*, Edizioni dell'Università degli studi di Cassino, Cassino, pp. 303-311.

Lupinu, G. (2003), *Tra latino epigrafico e sardoromanzo: sulla datazione di alcuni sviluppi fonetici*, in «Verbum», 5, 1, pp. 59-68.

Mancini, M. (2005), *La formazione del neostandard latino: il caso delle* differentiae uerborum, in Kiss, S., Mondin, L. and Salvi, G. (2005, *éds.*), *Latin et langues romanes. Études linguistiques offertes à J. Herman à l'occasion de son 80ème anniversaire*, Niemeyer, Tübingen, pp. 137-155.

Mancini, M. (2006), Dilatandis litteris*: uno studio su Cicerone e la pronunzia 'rustica'*, in Bombi, R., Cifoletti, G., Fusco, F., Innocente, L. and Oriolles, V. (2006, *a cura di*), *Studi linguistici in onore di Roberto Gusmani*, Edizioni dell'Orso, Alessandria, pp. 1023-1046.

Mannocci, A., Casarosa, V., Manghi, P. and Zoppi, P. (2017), *The* EAGLE *data aggregator: Data quality monitoring*, in Orlandi, S., Santucci, R., Mambrini, F. and Liuzzo, P.M. (2017, eds.), *Digital and Traditional Epigraphy in Context. Proceedings of the EAGLE 2016 International Conference*, Sapienza Università Editrice, Roma, pp. 161-172.

Marichal, R. (1992), *Les ostraca de Bu Njem*, Grande Jamahira Arabe, Libyenne, Populaire et Socialiste, Département des Antiquités, Tripoli.

Marotta, G. (2015), *Talking stones. Phonology in Latin inscriptions*, in «Studi e Saggi Linguistici», 53, 2, pp. 39-63.

Marotta, G. (2016), *Sociolinguistica storica ed epigrafia latina. Il corpus* CLaSSES I, in «Linguarum Varietas», 5, pp. 145-159.

Molinelli, P. (2006), *Per una sociolinguistica del latino*, in Arias Abellán, C. (2006, *éd.*), *Latin vulgaire - Latin tardif VII. Actes du VII^e^ colloque international sur le latin vulgaire et tardif*, Secretariado de Publicaciones Universidad de Sevilla, Sevilla, pp. 463-474.

Mommsen, Th. (1883, *Hrsg.*), *CIL X 1: Inscriptiones Bruttiorum, Lucaniae, Campaniae, Siciliae, Sardiniae latinae. Pars posterior inscriptiones Siciliae et Sardiniae comprehendens*, Reimer, Berlin.

Orlandi, S. (2017), *Il progetto* EAGLE. *Europeana network of Ancient Greek and Latin Epigraphy e le sue molteplici sfide*, in Mastandrea, P. (2017, *a cura di*), *Strumenti digitali e collaborativi per le Scienze dell'Antichità*, Edizioni Ca' Foscari, Venezia, pp. 49-60.

Orlandi, S., Santucci, R., Mambrini, F. and Liuzzo, P.M. (2017, eds.), *Digital and Traditional Epigraphy in Context. Proceedings of the* EAGLE *2016 International Conference*, Sapienza Università Editrice, Roma.

Panciera, S. (2013), *Notizie da* EAGLE, in «Epigraphica», 75, pp. 502-506.

Panciera, S. *et al.* (1991), *Epigrafia. Actes du Colloque international d'épigraphie latine en mémoire de Attilio Degrassi pour le centenaire de sa naissance*, Università di Roma 'La Sapienza' / École Française de Rome, Roma.

Poli, D. (1999), *Il latino tra formalizzazione e pluralità*, in Poccetti, P., Poli, D. and Santini, C. (1999, *a cura di*), *Una storia della lingua latina. Formazione, usi, comunicazione*, Carocci, Roma, pp. 377-431.

Prandoni, C., Casarosa, V. and Alfarano, N. (2014), EAGLE *Portal* [available online at *https://www.eagle-network.eu/wp-content/uploads/2013/06/EAGLE_D5.2_EAGLE-Portal_v1.0.pdf*].

Prandoni, C., Fresa, A., Alfarano, N., Amato, G., Zoppi, F., Mannocci, A., Liuzzo, P.M., Mambrini, F., Orlandi, S. and Santucci, R. (2017), *Searching inscriptions through the* EAGLE *Portal*, in Orlandi, S., Santucci, R., Mambrini, F. and Liuzzo, P.M. (2017, eds.), *Digital and Traditional Epigraphy in Context. Proceedings of the* EAGLE *2016 International Conference*, Sapienza Università Editrice, Roma, pp. 173-185.

Rocco, A. (2017), *EDB 2.0. How Eagle Europeana project improved the Epigraphic Database Bari*, in Orlandi, S., Santucci, R., Mambrini, F. and Liuzzo, P.M. (2017, eds.), *Digital and Traditional Epigraphy in Context. Proceedings of the* EAGLE *2016 International Conference*, Sapienza Università Editrice, Roma, pp. 115-130.

Rochette, B. (1997), *Le latin dans le monde grec*, Latomus, Bruxelles.

Rovai, F. (2015), *Notes on the inscriptions of Delos. The Greek transliteration of Latin names*, in «Studi e Saggi Linguistici», 53, 2, pp. 163-185.

Sotgiu, G. (1961), *Iscrizioni latine della Sardegna. Supplemento al Corpus Inscriptionum Latinarum, X e all'Ephemeris Epigraphica, VIII)*. Vol. 1, CEDAM, Padova.

Sotgiu, G. (1968), *Iscrizioni latine della Sardegna*. Vol. 2, 1: *Instrumentum domesticum. Lucerne*, CEDAM, Padova.

Sotgiu, G. (1988), *L'epigrafia latina in Sardegna dopo il C.I.L. X e l'E.E. VIII*, in Temporini, H. and Haase, W. (1988, *Hrsg.*), *Aufstieg und Niedergang der römischen Welt (*ANRW*)*. Vol. 2: *Principat*, 11.1, De Gruyter, Berlin / New York, pp. 552-739.

Tamponi, L. (2020), *Sardinian Latin through Inscriptions: A Variationist Analysis*, PhD thesis, University of Pisa.

Vineis, E. (1984), *Problemi di ricostruzione della fonologia del latino volgare*, in Vineis, E. (1984, *a cura di*), *Latino volgare, latino medioevale, lingue romanze*, Giardini, Pisa, pp. 45-62.

Vineis, E. (1993), *Preliminari per una storia (e una grammatica) del latino parlato*, in Stolz, F., Debrunner, A. and Schmidt, W.P. (1993, *a cura di*), *Storia della lingua latina*, Pàtron, Bologna, pp. xxxvii-lviii.

Wachter, R. (1987), *Altlateinische Inschriften. Sprachliche und epigraphische Untersuchungen zu den Dokumenten bis etwa 150 v. Chr.*, Peter Lang, Bern / Frankfurt am Main / New York / Paris.

Warmington, E.H. (1940), *Remains of Old Latin*. Vol. 4: *Archaic Inscriptions*, Harvard University Press / Heinemann, Cambridge, MA / London.

Giovanna Marotta
Dipartimento di Filologia, Letteratura e Linguistica
Università di Pisa
Via Santa Maria 36
56126 Pisa (Italy)
giovanna.marotta@unipi.it

Francesco Rovai
Dipartimento di Filologia, Letteratura e Linguistica
Università di Pisa
Via Santa Maria 36
56126 Pisa (Italy)
francesco.rovai@unipi.it

Irene De Felice
Dipartimento di Lingue e Culture Moderne
Università di Genova
Piazza Santa Sabina 2
16124 Genova (Italy)
irene.defelice@edu.unige.it

Lucia Tamponi
Dipartimento di Filologia, Letteratura e Linguistica
Università di Pisa
Via Santa Maria 36
56126 Pisa (Italy)
lucia.tamponi@fileli.unipi.it

SSL

Theoretical and pragmatic considerations on the lemmatization of non-standard Early Medieval Latin charters

Timo Korkiakangas

Abstract

This paper discusses the theoretical bases as well as the pragmatic implementation of the lemmatization of the Late Latin Charter Treebanks (*LLCT*). *LLCT* is a set of three dependency treebanks (*LLCT1*, *LLCT2*, *LLCT3*) of Early Medieval Latin documentary texts (charters) written in Italy between AD 714 and 1000 (c. 594,000 tokens). The original model for the lemmatization of *LLCT* was the Latin Dependency Treebank (*LDT*), which is mainly Classical standard Latin and based on the entries of Lewis and Short's *Latin Dictionary*. Since *LLCT* reflects later linguistic developments of Latin and contains a plethora of non-standard proper names, particular attention is paid to how non-standard lexemes are lemmatized systematically to make the lemmatization maximally usable. The theoretical underpinnings to manage the lemmatization boil down to two principles: the evolutionary principle and the parsimony principle.

Keywords: treebank, lemmatization, standardization, Medieval Latin charters, onomastics.

1. *Introduction*

Lemmatization: The reduction of the word tokens in a corpus to their lexemes. Thus, the set of word forms or tokens *swim*, *swam*, *swum*, *swims* and *swimming* constitute the lemma for the lexeme SWIM. 'Lemma' is mainly used as an alternative to 'lexeme' or 'headword', the form that heads an entry in a dictionary. (Brown and Miller, 2013: 259)

This paper interprets the above definitions in the way that lexemes are units of lexical meaning while lemma is the form of a lexeme which is conventionally chosen to represent the lexeme. In Latin, noun lemmas are presented in the masculine, neuter, or feminine nominative singular form, depending on the noun's gender; adjectives and pronouns are presented in the masculine nominative singular form. Verbs are given either the present

Received: February 2020
Accepted: April 2020

infinitive form or the first-person singular form of the indicative present. In *LLCT*, the latter form is chosen. With indeclinable parts of speech, the only form is the lemma. Latin lemmatization may look uncontroversial, but things become increasingly complicated as soon as concrete work begins, let alone with non-standard varieties of Latin.

There are currently no generally accepted guidelines for the lemmatization – or the morphological annotation – of Latin. In fact, no publication whatsoever exists that presents a set of principles sufficient for an exhaustive lemmatization or morphological annotation of Latin treebanks, hence the motivation of this special issue. On the one hand, this at first glance surprising defect is possibly motivated by the naïve image, probably fostered by unavoidably restricted normative school teaching, that Latin grammar is straightforward with its exhaustively described, well-defined grammatical categories and transparent lemmas. While this image is not completely distorted within the relatively narrow and well-codified linguistic landscape of Classical Latin, it is plainly untrue for any non-Classical, non-standard variety of Latin. On the other hand, the lack of lemmatization guidelines also seems to arise from the difficulty in systematizing the Latin lexicon satisfactorily, a task that should necessarily be based on extensive lexicographical work. The outcome has been that each project basically follows its own principles of lemmatization and morphological annotation. These principles are typically only described in passing, if at all, in publications on other topics (e.g. Philippart de Foy, 2012; Longrée and Poudat, 2010; McGillivray, 2014). The harmonization of the lemmatization between different Latin resources pursued within the Linking Latin (*LiLa*)[1] project at the Catholic University of Sacred Heart in Milan will no doubt help in establishing a solid ground on which to build a future consensus on Latin lemmatization.

The fluidness of the state of the art is also the reason why the lemmatization of *LLCT* does not form an integral whole. The lemmatization of *LLCT* is a hybrid of various usages adopted pragmatically and, to a certain degree, opportunistically from various sources, mainly from the Latin Dependency Treebank (*LDT*), and supplemented by *ad hoc* practices that looked adequate to manage given non-standard features of charter Latin. A special challenge of *LLCT* is the highly frequent proper names and especially the proper names of Germanic origin with no canonized spelling in Latin. Thus, the aim of this paper is to describe the principles followed in

[1] Cf. *https://lila-erc.eu/#page-top*.

the lemmatization of *LLCT* as exhaustively as possible. The discussion of the lemmatization principles will most often involve the *LLCT* treebanks as a whole (referred to as *LLCT*) while, occasionally, the focus will be on a single treebank (referred to as *LLCT1*, *LLCT2*, and *LLCT3*).

The discussion is organized as follows: Section 2 presents the *LLCT* treebanks while Section 3 briefly characterizes the type of Latin used in charters and defines what is meant by 'standard' in this paper. By giving some numerical data on lemmas in *LLCT*, Section 4 sets the background for Section 5, which discusses the two principles underlying the lemmatization of *LLCT*: the evolutionary principle (Section 5.1) and the parsimony principle (Section 5.2). Section 6 is the conclusion.

2. *The* LLCT *treebanks*

The *LLCT* treebanks consist of three morphologically and syntactically annotated corpora (*LLCT1*, *LLCT2*, *LLCT3*), which also feature a textual annotation layer that indicates abbreviated and restored words. Together the *LLCT* treebanks form a substantial resource for the research of the non-standard non-literary Latin of the Early Middle Ages[2]. Two of the *LLCT* treebanks (*LLCT1* and *LLCT2*) are thus far completed and openly accessible online[3]. The third part, *LLCT3*, is under construction and scheduled to be completed by 2021. *LLCT1* contains 225,834 tokens distributed within 519 charters written in Tuscany between AD 714 and 869, while *LLCT2* contains 257,819 tokens in 521 Tuscan charters from between AD 774 and 897. *LLCT3* will contain ca. 110,400 tokens in 221 charters written in Tuscany as well as in several locations in northern and southern Italy between AD 721 and 1000. The sources of *LLCT1* and *LLCT2* are five copyright-free editions published between 1833 and 1933: Barsocchini (1837), Barsocchini (1841), Bertini (1836), Brunetti (1833), Schiaparelli (1929), and Schiaparelli (1933a). Since most of the charters have also been published recently in the

[2] The other three Latin treebanks are the Latin Dependency Treebanks (*LDT*, *https://perseusdl.github.io/treebank_data/*), the *PROIEL* treebanks (*https://proiel.github.io*), and the Index Thomisticus Treebank (*IT-TB*, *https://itreebank.marginalia.it*).

[3] *LLCT1* is available in Prague Markup Language (*PML*) format at *https://zenodo.org/record/3633607#.XjU4lSNS9EY* and *LLCT2* in *CoNLL* format at *https://zenodo.org/record/3633614#.XjU6zCN7lEY* as well as in the *CoNLL-U* format on the website of the Universal Dependencies consortium at *https://github.com/UniversalDependencies/UD_Latin-LLCT/tree/dev* (see Cecchini *et al.*, 2020).

Chartae Latinae Antiquiores (*ChLA*) series, examples (1) to (6) of the present article will be conveniently referred to by their *ChLA* numbering. For a detailed description of the *LLCT* treebanks, see Korkiakangas (in press)[4].

The syntactic annotation of *LLCT* is based on dependency grammar as operationalized by the *Guidelines for the Syntactic Annotation of Latin Treebanks* (version 1.3; Bamman *et al*., 2007), which, for its part, complies with the annotation style adopted in the Prague Dependency Treebank (Hajič *et al*., 1999). Due to the above-discussed lack of generally accepted guidelines for the morphological annotation or lemmatization of Latin, the lemmatization and morphological annotation of *LLCT1* first practically imitated the choices made in the Latin Dependency Treebanks (*LDT*) available in 2010, the date of the first *LLCT* annotations. The *LDT* lemmas are derived from the Perseus Dynamic Lexicon, which is originally based on Lewis and Short's (1879) *Latin Dictionary* (Bamman and Crane, 2011: 11-13). *LLCT1* was lemmatized and annotated in the Perseus annotation environment, where the Dynamic Lexicon suggested possible lemmas when available. However, it soon became obvious that while the *LDT* style worked for the standard Latin forms of *LLCT*, both a considerable extension of the Perseus Dynamic Lexicon and a set of additional annotation rules were needed to manage the Early Medieval non-standard forms. These rules, described in Korkiakangas and Passarotti (2011), mostly specify principles related to the annotation of morphology, but they also briefly report decisions relative to lemmatization. The same lemmatization practice was originally used with *LLCT2*, which was automatedly annotated and then manually corrected.

The annotation and lemmatization of *LLCT2* were recently thoroughly revised prior to its conversion into the Universal Dependencies style[5]. In its present state, the lemmatization of *LLCT2* can no longer be identified with that of the *LDT* treebanks, based on the Perseus Dynamic Lexicon. At the same time, the possibility of making direct lemma-level comparisons with the *LDT* treebanks is lost. The current lemmatization of *LLCT2* represents a simplified version of the *LDT* style, independent of any predefined lexicon. This style is being utilized for the lemmatization of *LLCT3* as well. In comparison with the newly revised *LLCT2*, the annotation of *LLCT1* looks partly incoherent and should clearly be revised in the future.

[4] For various aspects of the morphological, syntactic, and textual annotation of *LLCT*, see Korkiakangas and Passarotti (2011) and Korkiakangas and Lassila (2013).

[5] The converted version will be distributed in a subsequent release of the Universal Dependencies at the project's website: *https://universaldependencies.org/#language-*.

3. *Early Medieval charter Latin*

Thousands of original Early Medieval charters survive in Italian archives. Charters are legal documents which record private transactions or trials. They were written by quill on parchment by professional or unprofessional lay or ecclesiastical scribes. Charters usually take up one parchment sheet and contain 200 to 1,000 words.

The language of legal documents is always formulaic, and Early Medieval charter formulae draw on a centuries-old legal Latin tradition. However, previous studies suggest that Early Medieval Italian scribes did not copy charters from formulary books, as was done later in the Middle Ages, but had memorized the conventional wordings which they then reproduced with varying success (Amelotti and Costamagna, 1975: 215-216; Schiaparelli, 1933b: 3), hence the considerable linguistic variation. In this way, features of the spoken language, which had evolved far from Classical Latin, occasionally ended up in Early Medieval Italian charters.

Because of this gap between the spoken and written codes, Early Medieval writers had to learn the written code of Latin practically as a second language (Korkiakangas, 2018: 441). Although the gap was wide, the *LLCT* charters suggest that it was still quantitative rather than qualitative. It looks likely that no meta-linguistic split was felt between the spoken language and its written form, both being still considered different sides of one language, Latin. Also, beyond the context of charters, a consciousness of two conceptually different languages seems to have emerged quite slowly in terms of written Latin and spoken Italo-Romance vernacular, a development that eventually led to the first attempts to establish a written form even for the latter (Wright, 2000). The first known reliably datable short texts in the vernacular date from the ninth and tenth centuries, but substantial texts only begin to appear in the following centuries (Frank-Job and Selig, 2016).

Given that Classical Latin standard had to be learnt, the departures from it could be held to be symptoms of the writers' poor school instruction. However, Bartoli Langeli (2006: 25), among others, maintains that, with all its spoken features, charter Latin had established itself as a cherished traditional Italian genre under the Lombard reign («national literature of Lombard Italy»). Be this as it may, charter Latin can be characterized as a 'non-standard' mixture of prefabricated formulae and spoken-language features, where archaic legal terminology is mingled with mistakes and hyper-

corrections provoked by the distance between the sought-after written code and the reality of the spoken language.

At this point, a definition of the term 'standard' (as an opposite of 'non-standard') is needed. In this paper, the 'standard' Latin of the Early Middle Ages refers to a Latin which essentially follows the spelling and morphology of Classical Latin as codified in the prescriptive grammars and used by the Christian authors of the Late Antiquity, who were considered models for literary activity throughout the Early Middle Ages. The spelling and morphology of the Latin of this type show only marginal deviations from those of the Classical Latin of the late Republic and the early Empire while more variation is observed in vocabulary and syntax. This type of standard grammar was still considered the model of written language in Tuscany of the eighth and ninth centuries, judging from other texts of the time as well as from the language of the best *LLCT* scribes. In sum, a rather clear point of reference in terms of a substantial consensus about 'correct' or 'accepted' language use was available in Early Medieval Italy (Korkiakangas, 2017: 577; Bartoli Langeli, 2006: 25 ff.)[6]. However, not all the scribes attained this standard, hence the notable inter-writer variation attested in *LLCT*.

4. *Overall description of the* LLCT1 *and* LLCT2 *lemmatization*

This section provides a background for the following sections by presenting a numerical panorama of the lemmatization of the two parts of *LLCT* already completed, *LLCT1* and *LLCT2*.

Table 1 shows that *LLCT1* contains 4,740 lemmas altogether. The lemma/token ratio is exceptionally low, only 2.1%, which means that each lemma is repeated around fifty times on average. This is because the most common formulae are repeated hundreds of times in the 521 charters of *LLCT1*. 2,139 of the lemmas were available in the Perseus Dynamic Lexicon while the remaining 2,601 lemmas, corresponding to 54.9% of all the lemmas, had to be added manually. 79.8% of the added lemmas were proper names; of all the *LLCT1* lemmas, proper names constitute 49.6%. Moreover, several proper name lemmas only appear once or a few times. These figures reflect well the special nature of charter Latin: many persons involved in the trans-

[6] Cf. Auernheimer's (2003: 49-51) decision to set Alcuin's (essentially Classical) Latin as the point of reference for her study on the Latin of the Carolingian hagiography.

actions are identified, whereas the text proper repeats the same wordings pertinent to its document type (e.g. lease, sales contract, donation) from charter to charter.

LLCT1			*LLCT2*		
tokens	225,834		tokens	257,819	
- lemmas	4,740		- lemmas	3,531	
- of which proper names	2,351	49.6%	- of which proper names	1,860	52.7%
- from *LDT*	2,139	45.1%	- from *LLCT1*	2,428	68.8%
- manually added lemmas	2,601	54.9%	- manually added lemmas	1,103	31.2%
- of which proper names	2,075	79.8%	- of which proper names	805	73.0%
lemma/token ratio	2.1%		lemma/token ratio	1.4%	

Table 1. *Tokens and lemmas in* LLCT1 *and* LLCT2[7].

The overall picture of *LLCT2* is similar to *LLCT1*, although the lemma/token ratio is even lower, 1.4%, with each lemma being repeated over seventy times on average. Such a narrowing is a symptom of the unification of documentary production in the early 9th century, from which the majority of the *LLCT2* charters date. Non-professionals were excluded from notarial practice and establishing chancery traditions entailed a stricter adherence to given formulae (Korkiakangas, 2017: 587; Costambeys, 2013: 246-248), hence the more limited lemma repertoire. *LLCT2* only contains 3,531 lemmas, 2,428 of which (68.8%) were directly transferred from *LLCT1* by way of a simple multi-replace script. For this reason, there is no immediate way to assess to what extent the lemmatization of *LLCT2* coincides with that of *LDT*.

Every corpus of Latin has to decide how to treat certain graphical conventions which change from edition to edition. In the lemmatization of *LLCT*, the character *j* is used before a vowel, whether it was written *j* or *i* in the source edition. Instead, *u* before a vowel is either *u* or *v* depending on the source edition. The *w* of the source editions, attested in words of Germanic origin, is treated inconsistently. In the text of *LLCT1*, it is kept *w* while, in *LLCT2*, it is rendered into the digraph *vu*. The lemmatization utilizes *w* consistently throughout *LLCT*. In *LLCT1*, the traditional Latin convention is followed to capitalize the lemmas that indicate months and calendar

[7] Note that the disambiguation numbers utilized in *LLCT1*, such as *1* in *nomen1* (see Section 5.2), were ignored when calculating the percentages.

terms, such as Kalends, while only proper name lemmas are capitalized in *LLCT2*. *LLCT3* will follow the practices observed in *LLCT2*.

It also needs to be mentioned that *LLCT* uses artificial tokens with no proper lemma to mark gaps in the text (*lacunae*). The artificial tokens are 556 in *LLCT1* and 461 in *LLCT2*. Thanks to the formulaicity of charters, the part of speech of a missing or fragmentary token can often be deduced quite reliably, even without certainty about the exact missing word. In such cases, an artificial placeholder token is created and lemmatized as 'missing^token' in *LLCT2*. For example, in the subscription formula *ego David filio* [Propn] *rogatus* [--] "I, David, son of [Propn], having been asked [--]", a generic [Propn] stands for the proper name expected in that context. It is lemmatized as 'missing^token'. Sometimes, a gap cannot be restored at all, as is the case with the last part of the above example. Then, the artificial placeholder token [--] is used and again lemmatized with 'missing^token'. *LLCT1* is more primitive in its treatment of artificial tokens, which are just marked with '[...]' or '[.....]' and left unlemmatized.

5. *Principles observed in the lemmatization of* LLCT2

The principles presented in the following sections work together in the lemmatization of *LLCT2* and are here separated from each other only for explanatory purposes. The evolutionary principle is presented in Section 5.1, which is further divided into five subsections 5.1.1 to 5.1.5 according to the type of the lemma. Section 5.2 discusses the parsimony principle.

5.1. *Evolutionary principle*

A fundamental principle governing the lemmatization of *LLCT* as well as its morphological annotation is the evolutionary principle which relates the language of *LLCT* to the Classical Latin standard, this latter being understood in the sense explained in Section 3. This principle is also the most distinctive feature of *LLCT* in comparison with treebanks of standard Latin. The evolutionary principle reduces the linguistic variants provoked by language evolution to their standard Latin ancestors. As regards morphological annotation, this reduction sometimes requires an identification of complicated processes which involve both phonological and morphological change in the inflectional ending, whereas with lemmatization, mainly

those evolutionary processes that affect the word stem are concerned. Because word-final inflectional morphemes are used to encode grammatical information in Latin, the evolutionary processes affecting word stems are phonological by nature, with the exception of changes in the number of syllables (see *cuntitigeris* etc. below). Since the challenges related to the lemmatization of proper names partly differ from those related to common names and other parts of speech, the following two sections discuss all other words than proper names, while sections 5.1.3 and 5.1.4 focus on proper names.

5.1.1. *Non-proper-name words with a standard Latin variant*

As regards morphology, the evolutionary reduction of Early Medieval forms to standard Latin forms can be exemplified by the prepositional phrase in (1), where *annus singulus* "every (single) years" is annotated as an accusative plural. This is because the ending *-us* is a typical evolutionary outcome of the standard Latin accusative plural *-os* following the closure of unstressed vowels (Väänänen, 1981: 36). The standard Latin accusative plural is *annos singulos* while the attested *annus singulus* could be misinterpreted, at first sight, as a homonym standard Latin nominative singular *annus singulus*. Obviously, the nominative does not go with a preposition:

(1) *per annus singulus* (*ChLA*[1], 1126)
"every year"

As stated above, with most lemmas it is enough to take phonological evolution into consideration because the morphological change manifests itself principally in inflectional endings. For example, the *LLCT* form *istio* (standard *aestivum*) is lemmatized under *aestivus* "summer-time" (adjective), *anfora* (standard *amphora*) under *amphora*, and *castangneto* (standard *castanetum*) under *castanetum* "chestnut grove". Note that this is done in spite of the fact that forms such as *anfora*, *castangneto*, *presunsere* (standard *praesumpserit*, lemmatized under *praesumo* "to venture"), or *prenda* (standard *prehendat*, lemmatized under *prehendo* "to take"), could very well be lemmatized under their modern Italian successors *anfora*, *castagneto*, *presumo*/*presumere*, and *prendo*/*prendere*, respectively. These fully Italo-Romance forms are likely to have already been in use in the spoken idiom of the time. In other words, the lemmatization of *LLCT* does not seek to describe any particular synchronic stage of Early Medieval Latin. If it did, it should reconstruct contemporary lemmas. That is, however, hardly possible, given the lack of

consensus about Early Medieval spoken Latin. Instead, the lemmatization of *LLCT* seeks to explicate and, subsequently, dissolve the diachronic distance between the attested forms and their standard Latin counterparts in the way that the Latin of *LLCT* is lemmatized as if it were standard Latin[8].

Morphological considerations come into question with lemmas where the stem has undergone alterations in syllabic structure, as is the case with *trentas* (standard *triginta* "thirty") or *poterent* (standard *possent* "they could"). The form *cuntitigeris* seems to be a reduplication inspired by the non-composite stem *tetig-* (standard *contigerit* "he/she may seize"). The evolutionary principle is, however, applied to them in the same way as it is applied to those infrequent cases where a change seems to have taken place in the word formation strategy between standard Latin and Early Medieval Latin: for example, *quattuorcentos* (standard *quadringentos*), lemmatized under *quadringenti* "four hundred" in *LLCT*.

5.1.2. *Non-proper-name words with no standard Latin variant*

The evolutionary principle is relatively easy to observe with Latin-based words discussed in the previous subsection while words that have no standard Latin variant turn out to be problematic. They are often spelled in several different ways, with no binding evidence in favour of one form rather than another. The great majority of the *LLCT* words with no ancestor in standard Latin are nouns, especially proper names (see Section 5.1.4). As for common nouns, words with no obvious standard variant are either loans from other languages, mainly Germanic ones, or Late Latin neologisms. The former include, among others, *sculdahis/sculdais*, a high official under the Lombard reign, *cafagium/gahagias/gahagium* "fenced estate", and *curte/curtis*, which derives from the Greek *khórtos* "courtyard", but seems to have no established Latin spelling. Based on the consultation of the Database of Latin Dictionaries (Brepols)[9] as well as on Nicoletta Francovich Onesti's studies (2000; 2002; 2010) on Germanic loans in Early Medieval Latin and following a careful scrutiny of the word's attestations in *LLCT*, a form that is most likely the common ancestor of the attested forms in terms of its frequency and/or (morpho)phonological features is set to be the lemma. It is either simply picked up among the attested forms or reconstructed if no attested

[8] In the same vein, the morphological annotation of *LLCT* can be used to observe how standard Latin categories are manifested in the Latin of *LLCT*.

[9] Cf. *https://about.brepolis.net/database-of-latin-dictionaries/*.

form seems to represent a (morpho)phonologically plausible ancestor form. In this way, the words above were assigned the lemmas *sculdahis*, *gahagium*, and *curtis*, respectively. As lexicon was not in the core of the projects under which *LLCT1* and *LLCT2* were built, not as much attention was paid to the Germanic words as would have been needed. Therefore, the outcome is often unsatisfactory and sometimes even erroneous in the light of evidence that has turned up during a later consultation of the above-mentioned dictionaries and studies.

Late Latin neologisms are more transparent than Germanic loans. Neologisms can often be assigned, with relative ease, a reconstructed lemma which complies with standard Latin morphology and spelling. This is particularly undisputed when neologisms are derived from standard Latin lexemes by way of usual word formation rules. For example, the adjective *massaricius* "pertinent to a villein holding" and the noun *massarius* "villein, tenant farmer" are regular Early Medieval derivations from the standard *massa* "parcel of land, villein holding" and can be adopted as standard Latin-like lemmas. The same applies to *mustariolum* "wine press", derived from *mustarius* "pertinent to must", or to *patrinius* "stepfather", cf. Italian *patrigno*, originally derived from *pater* "father". In the same vein, standard Latin-like lemmas are coined for less straightforward cases where the derivation involves no affixes and standard Latin models are less frequent: for example, the compound *modilocus* "area which yields one modius", derived from *modius* "corn measure" and *locus* "place, area" (Niermeyer *et al*., 2002, eds.: 911), *reddebeo* "to owe", derived from *reddo* "to pay" and *debeo* "to have to"[10], or the compound pronoun *tumetipse* "you yourself" for *temedipsa* in the phrase *per temedipsa* "by you yourself".

Finally, there are non-derived Early Medieval formations whose origin is not completely transparent: for example, *montone* "sheep" is lemmatized in *LLCT* under *monto*, which seems to be a variant of *multo* "mutton, sheep", cf. Old French *mutun*, modern French *mouton*. Likewise, *sellos* in *sex sellos de olibis* "six measures of olives" is lemmatized under *sellus*, a measure of capacity, possibly originally derived from *situlus* "bucket". If this interpretation is correct, the form postulates a development of the /tul/ group in /l:/ differently from the normal Italo-Romance pattern, where the regular phonological development resulted in /tul/ > /tl/ > /kl/ > /k:j/, like in modern Italian *secchio* (Väänänen, 1981: 65-66); cf. dialectal French *seille*, modern standard French

[10] Cf. Niermeyer *et al.* (2002, eds.: 1169) who use the lemma *redibere*, instead.

seau. Even the meaning of a word may remain unknown, as with *rasula* in the phrase *fini ipsa rasulam de bineam nostras* "up to the *rasula* of our vineyard"[11]. Nevertheless, the form is lemmatized under *rasula*. In this respect, the etymology principle is, in fact, typical of Romance linguistics, which routinely reconstructs 'proto-Romance' ancestors of Romance lexicon.

5.1.3. *Proper names of Latin origin*

As stated above, proper names pose particular challenges to lemmatization in *LLCT*. Since both anthroponyms and toponyms are particularly frequent in charters that record legal transactions between individuals at a certain place and time, a sound treatment of proper names is of the essence in *LLCT*. The challenges are related to two factors, the first of which is specific to *LLCT*: personal names of Germanic origin with no standard Latin ancestors were in fashion in Early Medieval Italy. The lemmatization of the names of Germanic origin involves a number of linguistic problems, which makes them the biggest stumbling block of *LLCT* lemmatization. The other reason is a global one: both anthroponyms and toponyms differ conceptually from common nouns in that their very form has a crucial informational function in identifying the language-external entity to which the name refers.

Proper names are subject to phonological change in the same way as all vocabulary of a given language, but because of their special informational function, they often tend not to be restored to their etymological standard forms in writing even when the writer might have known it, contrary to other vocabulary. As the semantic 'sense' of proper names is subordinate to their 'onymic', i.e. naming, reference (Anderson, 2007: 116 ff.), the etymological roots of names also become forgotten more readily than with normal vocabulary[12]. However, there seems to be a certain gradation in the maintenance of the form of names in *LLCT*, with names of particular importance or familiarity appearing more consistently in a form which was probably commonly felt to be the correct one and which sometimes also involved etymologization, especially if the name had standard Latin models. At least, the names of rulers and of the most important saints testify to such a tendency in *LLCT*, although even they vary quite a lot. On the other hand, the aspiration to restore names to their real or assumed standard Latin forms also

[11] The meaning "abrasion of skin" proposed in Du Cange *et al.* (1883-1887: *s.v. rasula*) does not make sense in this particular context where rather an agricultural term would be expected.

[12] For a detailed discussion on the special features of proper names, see Anderson (2007: § 4).

varies from writer to writer, with a few scribes preferring, for example, the hypercorrect *Latiarus* to *Lazarus* and *Austripertus* to *Ostripertus*.

In general, those proper names that have ancestors in standard Latin are lemmatized following the etymology principle as explained in Section 5.1.1. This is uncontroversial in transparent cases, such as *Pretestatus* (lemmatized under *Praetextatus*), *Deusdede* (lemmatized under *Deusdedit*), originally Greek *Aeleutieri* (lemmatized under *Eleutherius*), or toponym *Ilice* (lemmatized under *Ilex*). However, it is sometimes difficult to decide whether certain names, such as *Liliodarus*/*Lilioderus* or *Theopingtus*/*Thepingtus*, originally have ancestors in standard Latin or whether they are rather combinations of Latin and Germanic elements, like, for example, *Clarisinda* clearly seems to be. *Liliodarus* and *Lilioderus* are lemmatized under *Liliodorus* and may be originally composed of *lilium* "lily" and *dōron* "gift", a typical element of Greek anthroponyms. *Lilio-* is also attested in other *LLCT* names, such as *Liliaufunsus* (lemmatized under *Liliofonsus*), *Liliopinctus*, and *Liliolus*. *Theopingtus* and *Thepingtus* are lemmatized under *Theopinctus*. On the one hand, the name could be a variant of the late Greek *Theópemptos* or *Theópentos* while, on the other, *pinctus* may mean "decorated, adorned", from *pingo* "to paint", a meaning that would make sense in *Liliopinctus*; cf. Italian compounds, such as *variopinto* "multicolour". The first element of *Theopingtus*/*Thepingtus* can also be inspired by Germanic names, such as *Teutfrid* and *Teopaldo*, which begin with the popular Germanic element *t*(*h*)*eu-*/*t*(*h*)*eo-* (< **þeuð-* "tribe, people") (Francovich Onesti, 2000: 216; Francovich Onesti, 2002: 1142).

With some undoubtedly Latin-based names, it is not obvious what the original form is, as phonological development has obscured it and several close variants may occur side by side. This situation is typical of toponyms. For example, it can be duly asked whether the forms *Rocta*, *Ropta*, *Rotta*, and *Rota* are different spelling variants of the same toponym. The first three quite likely derive from the standard Latin participle *rupta* "broken, i.e. rocky", while the last one could equally well come from *rota* "wheel". Based on topographical considerations, they are all lemmatized under *Rupta*.

Any uncertainty about the standard Latin ancestor form of names that only occur in one form in *LLCT* leads to the sole attested form being taken up as the lemma: for example, the toponym *Coltserra* or the anthroponym *Inquircius*. As the *LLCT* treebanks were lemmatized over a long period of time, new instances kept turning up over the process that called for a reappraisal of the previously assigned lemma. The lemmatization has sometimes

failed to be changed accordingly, a fact that contributes to the present incoherent state of the lemmatization of proper names in *LLCT*. Moreover, a deliberate differentiation is sometimes applied in cases where there is insufficient proof to identify two or more slightly differently spelled anthroponyms or toponyms with each other. For example, it is not sure that *Sarturiano* and *Satoiano* (lemmatized under *Sartorianum* and *Satoianum*, respectively) refer to the same place even though that seems possible on phonological grounds. All this having been said, there is no doubt that a scrupulous onomastic revision would radically improve the lemmatization of *LLCT*. As mentioned above, the reason behind the present deficiencies in the lemmatization of proper names is that onomastics did not rank among the interests that guided the building of the *LLCT* treebanks, where the focus has always been on morphology and syntax rather than vocabulary.

Sometimes, it is not clear whether a second-declension toponym that ends in *-o* should be interpreted as neuter or masculine. This is because the neuter as an independent gender category had practically disappeared by the Early Middle Ages and because the *-o* ending can be argued to represent the Romance-type default form of the singular *-o* declension, derived from the accusative in *-u*(*m*) (for both masculine and neuter; Smith, 2011: 278, with references; Korkiakangas, 2016a: 291-295; Korkiakangas, 2016b: 72-73). It was decided that with toponyms ending in *-o*, the *LLCT* lemma ends in *-um* if it is not clearly based on a certain unquestionably reconstruable form of other gender, as is the case with *Saltucclo*, which must be derived from the masculine noun **saltuculus* (diminutive of *saltus* "forest") and is lemmatized as such (*Saltuculus*). For example, the toponym *Sexto* (modern *Sesto*) in *de loco Sexto* "of the place Sexto" and in *ad Sesto* is lemmatized under the neuter noun *Sextum*, although it could also be lemmatized under the masculine adjective *Sextus*, especially when it occurs with *loco* "place". However, in most cases, the elliptical *loco* construction cannot be used as a proof because it allows lack of agreement: for example, the feminine noun in *in loco Valeriana* and the genitive in *in loco Capelle*. Regrettably, an opposite decision was made concerning those third-declension toponyms whose gender cannot be deduced from the form attested in *LLCT*, such as *Lunise* in *ad Lunise* or *Montise* in *ubi dicitur Montise* "which is called Montise". They were interpreted as masculine accusative forms and assigned the masculine lemmas *Lunensis* and *Montensis*, respectively, despite the fact that the forms in question could be neuter (or feminine) accusatives as well. Third-declension toponyms of this kind are infrequent, though.

5.1.4. *Proper names of Germanic origin*

As was suggested above, the lemmatization of proper names of Germanic origin is even less accurate and less coherent than that of Latin-based (or originally Greek-based) names. Therefore, it is not recommended to use the lemmatization of *LLCT* for onomastic investigations.

The evolutionary principle cannot usually be sensibly applied to the Germanic names that occur in *LLCT* because they almost never have obvious standard variants. The cases closest to a standardization of any kind include rulers' names, such as *Carolus*/*Karolus* or *Berengario*, lemmatized under *Carolus* and *Berengarius*, respectively. As a rule, each name has to be evaluated separately based on research on historical Germanic languages. In this respect, the studies of Francovich Onesti (2000; 2002; 2010) have again been of great help, but, as stated above, they were not consulted in a systematic way under the construction phase of the *LLCT* treebanks. Moreover, knowledge on original Germanic morphological elements only helps in recognizing them behind Early Medieval Latin names and, thus, in unifying the spelling of that element in the lemmatization. Occasionally, it also helps in matching two very differently spelled names under one lemma.

However, Germanic morphology results in highly varying outcomes in the Latin of charters. For example, according to Francovich Onesti (2000: 173), the element **agjō* "blade" can be recognized in charters behind the elements *Agi-*, *Aghy-*, *Age-*, *Atge-*, *Ag-*, *Agg-*, *Agel-*, *Agil-*, *Achi-*, *Ahci-*, *Aci-*, *Ace-*, *Ac-*, *Acu-*, and *Ai-*. Yet, some Germanic-based onomastic elements seem to represent established Tuscan types: for example, the spellings *Achi-* and *Agi-* are particularly frequent in *LLCT*. Thus, even though it might be possible in some cases, it is of no use to seek to reduce the immense spelling variation conditioned by Early Medieval Latin phonology to any artificial Germanic lemma by creating lemmas beginning with *Agjo-* for this specific morpheme (e.g. *Agipert* lemmatized under fictitious *Agjoberhtaz*). Instead, it is possible to recognize whether a certain linguistically plausible form is clearly a preferred one in terms of frequency and then to use it as the lemma. Alternatively, the considerations on frequency and Germanic morphology may help reconstruct a lemma as the common denominator to all the attested forms. In spite of this, decisions have been difficult, and, for example, the forms *Agiulo*/*Aggioli* (genitive), *Agguli* (genitive), *Aculo*, and *Aiuli* (genitive) have ended up with four lemmas in *LLCT*, *Agiolus*, *Aggulus*, *Aculus*, and *Aiolus*, respectively, although there seems to be no reason not to consider them representatives of the same lemma, whatever that might be (perhaps *Agiolus*). Although the ap-

plication of the etymology principle is reduced with Germanic names, special care was taken to ensure that names that refer to a certain person are always lemmatized under one lemma. For example, *Hluttarius*, *Hlotharii* (genitive), and *Lotharii* (genitive), all referring to the king Lothar, are lemmatized under *Hlotharius*. The same applies to notaries or other identifiable persons that occur a number of times in one or in several charters. Further, lemmatization is sometimes inconsistent between *LLCT1* and *LLCT2*: for example, *Ildicari* (genitive) and *Ildechieri* (genitive) have mistakenly ended up with two lemmas, *Ildicarus* in *LLCT1* and *Ildecherus* in *LLCT2*.

The Germanic-based masculine names of *LLCT* appear either with Latin inflectional endings, with the Germanic ending *-i* (Francovich Onesti, 2000: 233), or without inflectional endings at all: for example, *Gunfridus*, *Gunfridi*, and *Gunfrid* are all attested. The choice between the three seems to be idiosyncratic, but the Latin endings are by far the most frequent. All the feminine names end in *-a* in the nominative singular (e.g. *Aliperga*) while the names that have entered the Latin third declension (e.g. *Frido*) are usually inflected according to the nasal paradigm (e.g. *Friduni*, dative; Francovich Onesti, 2000: 240) and are, consequently, easy to lemmatize (*Frido*). The *LLCT* lemmatization adds the Latin inflectional ending *-us* to those names that have entered the Latin second declension at least once in *LLCT*; for example, the above *Gunfridus*, *Gunfridi*, and *Gunfrid* are lemmatized under *Gumfridus*. Quite rare Germanic names, such as *Aloin*/*Aloni* or *Eoin*, never appear inflected in *LLCT*, hence their lemmatization without inflectional endings (*Aloin* and *Eoin*, respectively). This practice is identical with the one observed with Biblical names that are traditionally used uninflected and are lemmatized accordingly (e.g. *Daniel*, *Abraham*). Yet other names fluctuate between the second and third Latin declensions, which has sometimes led to inconsistent lemmatization decisions: the nominative and genitive form *Waltari* gets an accusative form *Uualtarene* and is lemmatized under *Waltarus*, although the genitive *Waltari* does not necessarily entail belonging to the second declension.

5.1.5. *Evolutionary principle with mistaken expressions*

This section discusses the import of the evolutionary principle on the lemmatization of mistaken words in *LLCT*. Such a scenario is irrelevant with literary corpora, where erroneous forms are not present, but is pertinent with charters, which are unemended original documents and feature significant linguistic irregularities.

Let us first consider how the evolutionary principle is applied to erroneous morphosyntax. In order to cope with the non-standard morphology of *LLCT*, Korkiakangas and Passarotti (2011: 106 ff.) coined an annotation principle based on 'functional' and 'formal' analyses of morphosyntax. The principle operates on the syntax/semantics interface, linking attested morphological forms to their standard Latin ancestors with the help of the evolutionary principle. Importantly, it also deals with erroneous forms that are impossible from the viewpoint of language evolution, i.e. motivated extra-linguistically. In such cases, an attested morphological form does not match with its expected standard Latin function on the syntax/semantics interface. For example, in (2), the coordinated ablative/dative form subject *heredibus nostris* "our heirs" depends on the predicate *habitare debeamus* "have to dwell".

(2) *Tam nos quam et heredibus nostris in ipsa casa habitare debeamus.* (*ChLA*[1], 1061)
"Both we and our heirs have to dwell in that house."

In standard Latin, the subject of the finite verb is always marked with the nominative case. The form *heredibus nostris* cannot be a morphophonological evolutionary outcome of the standard Latin nominative form *heredes*, and, therefore, it cannot be marked functionally as a nominative. *Heredibus nostris* must be a linguistic error due to a contamination between two or more formulae, a phenomenon frequent in charters, or to an infelicitous interpretation of the abbreviation *hhd* (for *heredes*) (Korkiakangas and Passarotti, 2011: 107). In *LLCT*, functionally impossible mistaken forms of this kind are simply assigned a formal morphological analysis that corresponds to the evolutionary ancestor of that form in standard Latin. Thus, *heredibus nostris* receives an ablative/dative plural morph tag although the subjects of finite verbs cannot be marked with such a case in any variety of Latin.

While the practice described above is fundamental to the annotation of non-standard morphology, it also plays a marginal role in lemmatization, where the question is basically about semantics. Words that are incongruent, i.e. mistaken, in their present context are found sporadically in *LLCT*. In literary texts, one is not accustomed to find mistaken words because literary texts are transmitted through centuries of copying and emendation and finally subjected to editing based on textual criticism. Instead, the scribes who wrote charters were not always equal to their tasks in this respect. Some

misunderstood expressions, usually in age-old documentary formulae, are characteristic of a single scribe, while others are used by more scribes, suggesting thus a local convention. For example, a few scribes mistakenly use in (3) the form *genium*, which looks like an accusative singular form of the word *genius* "tutelar deity, genius", in lieu of *ingenium* "natural disposition, machination, scheme"[13], a word that normally appears in the formula of (3) and makes sense in that context. The translation of (3) conveys the intended meaning (*ingenium*).

(3) *Si forsitans quicumquem de heredis meis* [...] *substraheret quesieret per colive genium*. (*ChLA*[1], 1058)
"If anyone of my heirs [...] perchance tries to dispossess [something] by whatever scheme."

Genium is not an evolutionary outcome of any morphophonological process of *ingenium*, but a blatant misinterpretation resulting from the writer having confused *ingenium* with *genius*, the latter most likely absent in the spoken vernacular of the time. In (3), *genium* is lemmatized under *genius*, which is the only possible standard Latin source for the attested form. This kind of lemmatization follows the practice of formal analysis observed with non-standard morphology and illustrates the uncompromising mode of operation of the evolutionary principle: it always reduces an attested form to its morphophonologically possible language-evolutionary ancestor, whether it makes sense or not in terms of the integrity of the construction or its meaning.

As stated, clearly mistaken words are relatively infrequent in *LLCT*. Additionally, with most mistakes, the formal analysis is obvious and the application of the etymology principle banal: this is the case if the attested word is completely different from the expected/intended one, such as *tradedimus* "(we) handed over/commissioned" in (4), where *rogavimus* "(we) asked" would have been expected on the basis of numerous occurrences. The translation again conveys the intended meaning (*rogavimus*).

(4) *Quam biro cartolas binditionis nostres ad nus factas Warnegausu notarium iscriberes tradedimus*. (*ChLA*[1], 732)
"We asked the notary Warnegausu to write these sales contracts which we made."

[13] Du Cange *et al.* (1883-1887: *s.v. ingenium*).

Tradedimus is not etymologically derived from *rogavimus*, which normally appears in this formula, and is lemmatized formally under *trado* "to hand over/to commission". The writer has probably confused the construction with *trado* with a gerund, which is, however, only attested once in charters (sentence in (5)). Here, the gerund is *scriuendo* "to be written" while the sentence in (4) shows an infinitive (*iscriberes*, i.e. *scribere*).

(5) *Ego Uualtprand in Dei nomine episcopus in hanc cartula donationis* [...] *manus meas suscribsi et confirma et scriuendo tradedi.* (*ChLA*[1], 911)
"I, Waltprand, bishop in God's name, subscribed [...] in this donation and confirmed [it] and commissioned [it] to be written."

In conclusion, it must be stated that the lemmatization of mistaken expressions in *LLCT* has not been as systematic as would be desired. In the sentence in (6), the writer has written *insunt* "(they) are in" instead of *hi sunt* "these are". The former is a nonsensical misinterpretation of the latter, which is the normal way to introduce a list of names in the formula in question and a variant of the frequent *id est* "i.e.". However, when the sentence was lemmatized for *LLCT2*, *insunt* was 'normalized' by splitting it into two tokens, and *in* lemmatized as *hi* "these" under *hic* "this" and *sunt* "(they) are" under *sum* "to be".

(6) *Direxistis missos tuos, in sunt Petrus notario de Uuamo et Sicholfo.* (*ChLA*[2], 85, 37)
"You sent your envoys, they are Petrus, the notary from Guamo, and Sicholfo."

That the writer has written *insunt* intentionally is proved by the following sentence, which lists another set of envoys and also features *insunt*. Thus, to be consistent with the practice described in this section and to respect the choices made by the charter scribes, *insunt* should be restored as one token and lemmatized under *insum* "to be in" as soon as *LLCT2* is again revised some day in the future.

5.2. *The Parsimony principle and homonymous lemmas*

The other general principle that is applied to the lemmatization of *LLCT* together with the evolutionary principle can be called 'the parsimony

principle'. This means that the lemmatization style of *LLCT* does not seek to multiply lemmas unnecessarily. As stated above, not only spelling, but also inflectional morphology fluctuated in Early Medieval Latin. One solution to cope with forms that have changed their inflectional properties is to provide these non-standard forms with new lemmas. This is what some dictionaries do when they provide separate entries to pre-Classical gender variants, such as *corium* (neuter) as opposed to *corius* (masculine) "skin" (e.g. Forcellini *et al.*, 1858-1875; Gaffiot, 1934). Such a solution does not, however, do justice to later written Latin, where borders between declensions, conjugations, and genders had become increasingly permeable in several morphophonological contexts (Sornicola, 2017: 85 ff.), without implying a change in meaning. Due to this inflectional flexibility, there is no reason to postulate new Early Medieval lemmas underlying the non-standard forms (Philippart de Foy, 2012).

Therefore, in *LLCT*, the new second-declension adjective *inanus* "void" (possibly reinforced by the second-declension *nanus* "dwarf", given that the form *nanis* is attested seven times in *LLCT1*) and the third-declension genitive/dative anthroponym *Ursoni* (genitive) "Ursus" with a Late Latin nasal declension are lemmatized under the corresponding standard lemmas: the third-declension *inanis* and the second-declension *Ursus*, respectively. This is done even though the ending *-us* is not etymologically derived from *-is* nor *-oni* from the standard genitive ending *-i*. A major subgroup of *LLCT* words with non-standard inflectional properties is formed by nouns which have undergone a gender change, such as *seculi* "centuries" in *super isti futuri seculi* "over the future centuries", where *seculi* with the masculine nominative plural ending *-i* is lemmatized under the standard Latin neuter *saeculum* (whose nominative plural is *saecula*) (Korkiakangas and Passarotti, 2011: 108). Likewise, *offertas* "offerings", seemingly a feminine accusative plural that had developed from the collective neuter plural in *-a*, *offerta*, is lemmatized under the Late Latin neuter singular lemma *offertum* (Adams, 2013: 431-432; Väänänen, 1981: 101-105).

An assignment of separate lemmas, such as *inanis* and *inanus*, to the above-mentioned standard and non-standard forms, respectively, would be a bad solution, not only because it ignores the historical development of Latin, but also because lemmas are by definition independent units of meaning, as stated in Section 1. In the cases above, inflectional change does not affect meaning. On the other hand, there are genuinely homonymous lexemes with different meanings that have to be lemmatized under separate lemmas

(Murphy, 2010: 84). An example of homonymous lemmas in English are (*to*) *lie* "(to) speak falsely" and (*to*) *lie* "(to) rest horizontally". They are sometimes registered under separate entries in English dictionaries, especially if they belong to different parts of speech, such as the above verbs and the noun *lie* "false statement".

In Latin, verb lemmas are rarely homonymous with lemmas of other parts of speech, contrary to English. Homonymous lemmas are potentially problematic for corpus linguistics, but in practice they are almost always disambiguated by their part of speech and syntactic properties. For example, the verb *intro* is inflected in person, tense, mood, and voice while the Late Latin preposition *intro* is indeclinable, and they have completely different distributions. An insuperable ambiguity only arises with lemmas such as *jus* "justice" vs *jus* "broth, juice", which are both nouns.

With *LLCT*, it is defined that homonymy arises when identical lemmas have different parts of speech or when they are etymologically of different origin. On the other hand, the evolutionary principle presented above entails that semantic differentiation does not give rise to new lemmas (Murphy, 2010: 87-90). For example, *band* "strip or loop of material" and *band* "musical group" in English would not be considered different lemmas in *LLCT* because they derive etymologically from the same origin. An opposite approach is seen, for example, in the *Longman Dictionary of the English Language* (Gay *et al*., 1984, eds.: 111), which gives the above nouns independent lemmas *[1]band* and *[3]band*.

Technically, the lemmatization of *LLCT1* follows the original *LDT* style in that homonymous lemmas are disambiguated by specifier numbers, for example *intro1* and *intro2*, with non-homonymous lemmas marked with *1* by default (e.g. *nomen1*). As was stated, the *LDT* lemmatization is based on the Perseus Dynamic Lexicon, which reproduces the entries of Lewis and Short (1879). Since Lewis and Short did not aim at keeping the lemmas at a minimum, the *LDT* style includes quite a number of cases with homonymous lemmas that could be subsumed under one lemma (e.g. *pecus* "cattle, beast" with three entries). Moreover, no clear distinction is made between past participles and homonymous nouns derived from them, such as *exitus* "gone out" and *exitus* "departure". Even *LLCT2* initially exploited specifier numbers, but in the current revised version of *LLCT2*, the numbers were removed and the ten remaining pairs of homonymous lemmas disambiguated by way of a specifier that usually indicates the part of speech, such as *latus^n*(*oun*) "side, flank" as opposed to *latus^a*(*djective*) "wide" and

intro^v(*erb*) as opposed to *intro^p*(*reposition*)[14]. This was done to respect the definition of a lemma as a semantically distinct unit, although in the case of *LLCT*, the use of specifiers is strictly speaking redundant, given that all the homonymous lemmas of *LLCT* can also be disambiguated by referring to the part-of-speech annotation layer. For the present, there are no genuinely homonymous lemmas in *LLCT*, such as the two nouns *jus*.

Having said all this, some borderline cases still remain in the lemmatization of *LLCT*. The word *locus* "place", originally a masculine, is very often used with the neuter endings *locum* and *loca*. The current version of *LLCT1* still lemmatizes forms with undeniable neuter endings under *locum1*, while the forms with endings that can be attributed to the masculine lemma go under *locus1*, contrary to the parsimony principle. In *LLCT2*, this incoherence has been corrected, and all forms are now lemmatized under *locus*. Likewise, in their current state, both *LLCT1* and *LLCT2* separate the lemmas *dominus* and *domnus*, although the latter clearly derives from the former. The lemma *dominus* "Lord" almost exclusively refers to God, while *domnus* "lord" is used as an appellation of human beings, e.g. *domnus Iacobus episcopus* "lord Jacobus, the bishop" (cf. Italian *don*). The treatment of *locus* has to be rectified in *LLCT1* and that of *domnus*/*dominus* both in *LLCT1* and *LLCT2* in pursuance of an anticipated general revision of *LLCT1*.

6. *Conclusion*

This paper has analysed the theoretical bases of the lemmatization of the Late Latin Charter Treebanks by discussing in detail the principles that were followed in their lemmatization: the evolutionary principle and the parsimony principle. In addition to the fact that no generally accepted guidelines for the lemmatization of Latin exist, the non-standard Early Medieval fea-

[14] The other homonymous lemmas marked with a specifier in *LLCT2* are *amicus^n*(*oun*) "friend" as opposed to *amicus^a*(*djective*) "friendly", not present in *LLCT*; *excepto^adv*(*erb*) "except" as opposed to *excepto^c*(*onjunction*) "except", *excepto^p*(*preposition*) "except (for)", and *excepto^v*(*erb*) "to exclude"; *finis^p*(*reposition*) "up to" as opposed to *finis^n*(*oun*) "end, region"; *intrinsecus^n*(*oun*) "indoor movables" as opposed to *intrinsecus^adv*(*erb*) "inwardly", not present in *LLCT*; *labor^n*(*oun*) "work" as opposed to *labor^v*(*erb*) "to glide", not present in *LLCT*; *papa^father* "father, pope" as opposed to *papa^pappa* "the word with which infants call for food" (Lewis and Short, 1879: *s.v. papa*), not present in *LLCT*; *partio^n*(*oun*) "part, portion" as opposed to *partio^v*(*erb*) "to share", not present in *LLCT* (*partio* may be a contamination of *portio* "portion" and *parte*(*m*) "part" or *partitio* "partition"); *super^p*(*reposition*) "over, above" as opposed to *super^adv*(*erb*) "over, above".

tures of charter Latin pose challenges to all levels of linguistic analysis, not least to lemmatization. Particularly, the highly frequent proper names with no canonized spelling in Latin are difficult to lemmatize consistently. Many of the most challenging names are of Germanic origin.

The central problem of the Latin of *LLCT* is how to use the analytical apparatus arising from Classical standard Latin to annotate forms and lemmatize words that do not exist in that standard. Because Early Medieval Latin never formed a written standard of its own, no description of its grammatical categories or its vocabulary is sufficiently solid to serve as the basis of morphological annotation or lemmatization, hence the adherence to the grammatical description of Classical standard Latin. In order to leap the gap between the attested non-standard forms and the existing standard, a principle called 'the evolutionary principle' was introduced. This principle reduces the linguistic variants provoked by language evolution to their standard Latin ancestors.

It is relatively easy to apply the evolutionary principle to Latin-based common names and other parts of speech which do have a standard Latin ancestor, while the lemmatization of forms that have no standard-Latin ancestor is more challenging. These latter are Late Latin neologisms or loans from other languages, mainly from Germanic ones, and they usually display a number of different spellings. The word's attestations in *LLCT* and in other sources, if available, are first carefully analysed and relevant lexicographical studies consulted. Subsequently, the (morpho)phonologically most plausible ancestor is either chosen between the attested forms or reconstructed on their basis.

Due to their special role in naming individuals, proper names tend to show more phonological erosion and less corrective normalization than other vocabulary and, therefore, their etymological origins become more readily blurred. This issue is pronounced in charters, where both anthroponyms and toponyms are frequent. Proper names with standard Latin ancestors are usually lemmatized with little uncertainty, while proper names with foreign, mainly Germanic, origin pose the biggest challenges to the use of the evolutionary principle: the Germanic names of *LLCT* almost never have obvious standard variants. The decision on the lemma is based on the frequency and the language-historical plausibility of the form. However, the *LLCT* lemma of a Germanic-based proper name is not a faithful reconstruction of the underlying Germanic word but rather an abstraction based on the attested Early Medieval Latin forms.

As charters are original documents and their Latin is highly irregular, the lemmatization, as well as the morphological and syntactic annotation, also have to take mistaken expressions into consideration. According to the evolutionary principle, functionally nonsensical semantic mistakes are not corrected in the lemmatization, just like functionally impossible mistaken morphology is annotated formally as it stands.

The other general principle applied to the lemmatization of *LLCT*, i.e. the parsimony principle, is introduced to avoid unnecessary proliferation of lemmas. The parsimony principle lumps under one lemma the forms that have the same meaning but have changed their inflectional properties. On the other hand, there are genuinely homonymous lexemes with different meanings that have to be lemmatized under distinct lemmas. Based on the evolutionary principle, identical lemmas are only considered homonymous in *LLCT* if they have different parts of speech and they are not of the same origin etymologically.

The scrupulous analysis of the above issues has shown that the lemmatization of *LLCT* is not as coherent as it should be. While the bulk of the lemmas of common nouns and other parts of speech can be trusted, the lemmatization of proper names would clearly benefit from a careful harmonization, hopefully realized in pursuance of a future revision of *LLCT1* and later a revision of *LLCT2*.

References

Adams, J.N. (2013), *Social Variation and the Latin Language*, Cambridge University Press, Cambridge.

Amelotti, M. and Costamagna, G. (1975), *Alle origini del notariato italiano*, Giuffrè, Milano.

Anderson, J. (2007), *The Grammar of Names*, Oxford University Press, Oxford.

Auernheimer, B. (2003), *Die Sprachplanung der karolingischen Bildungsreform im Spiegel von Heiligenviten*, K.G. Saur, München / Leipzig.

Bamman, D. and Crane, G. (2011), *The Ancient Greek and Latin Dependency Treebanks*, in Sporleder, C., van den Bosch, A. and Zervanou, K. (2011, eds.), *Language Technology for Cultural Heritage*, Springer, Berlin / Heidelberg, pp. 79-98.

Bamman, D., Passarotti, M., Crane, G. and Raynaud, S. (2007), *Guidelines for the Syntactic Annotation of Latin Treebanks*, v. 1.3. [available online at *nlp.perseus.tufts.edu/syntax/treebank/ldt/1.5/docs/guidelines.pdf*].

Barsocchini, D. (1837), *Memorie e documenti per servire all'istoria del Ducato di Lucca*. Tomo 5, 2, Francesco Bertini, Lucca.

Barsocchini, D. (1841), *Memorie e documenti per servire all'istoria del Ducato di Lucca*. Tomo 5, 3, Francesco Bertini, Lucca.

Bartoli Langeli, A. (2006), *Notai: scrivere documenti nell'Italia medievale*, Viella, Roma.

Bertini, D. (1836), *Memorie e documenti per servire all'istoria del Ducato di Lucca*. Tomo 4, 2, Francesco Bertini, Lucca.

Brown, K. and Miller, J.E. (2013), *The Cambridge Dictionary of Linguistics*, Cambridge University Press, Cambridge / New York.

Cecchini, F.M., Korkiakangas, T. and Passarotti, M. (2020), *A new Latin treebank for Universal Dependencies: Charters between Ancient Latin and Romance languages*, in Calzolari, N., Béchet, F., Blache, Ph., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J. and Piperidis, S. (2020, eds.), *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (*LREC *2020)*, European Language Resources Association (ELRA), Paris, pp. 933-942.

ChLA^{1} = Chartae Latinae Antiquiores*: Facsimile-Edition of the Latin Charters Prior to the Ninth Century*, Bruckner, A., Marichal, R. *et al.* (1954-2001, eds.), Urs Graf Verlag, Olten / Dietikon / Zürich.

ChLA^{2} = Chartae Latinae Antiquiores*: Facsimile-Edition of the Latin Charters, 2nd Series: Ninth Century*, Cavallo, G., Nicolaj, G. *et al.* (1997-2019, eds.), Urs Graf Verlag, Dietikon / Zürich.

Costambeys, M. (2013), *The laity, the clergy, the scribes and their archives: The documentary record of eighth and ninth-century Italy*, in Brown, W., Costambeys, M., Innes, M. and Kosto, A. (2013, eds.), *Documentary Culture and the Laity in the Early Middle Ages*, Cambridge University Press, Cambridge, pp. 231-258.

Du Cange, Ch., Charpentier, D.P., Henschel, G.A.L. and Favre, L. (1883-1887, [1678^{1}]), *Glossarium mediae et infimae latinitatis* (édition augmentée), Niort, Paris.

Forcellini, E., Furlanetto, G. and De-Vit, V. (1858-1875), *Totius Latinitatis Lexicon*. Voll. 1-4, Typis Aldinianis, Pratii.

Francovich Onesti, N. (2000), *Vestigia longobarde in Italia (568-774): lessico e antroponimia*, Artemide edizioni, Roma.

Francovich Onesti, N. (2002), *The Lombard names of Early Medieval Tuscany*, in Boullón Agrelo, A.I. (2002, *ed.*), *Actas do XX Congreso Internacional de Ciencias Onomásticas*, Fundación Pedro Barrié de la Maza, A Coruña, pp. 1141-1164.

Francovich Onesti, N. (2010), *Indizi di sviluppi romanzi riflessi nelle voci germaniche e nei nomi propri*, in «Germanic Philology», 2, pp. 67-101.

Frank-Job, B. and Selig, M. (2016), *Early evidence and sources*, in Ledgeway, A. and Maiden, M. (2016, eds.), *The Oxford Guide to the Romance Languages*, Oxford University Press, Oxford, pp. 24-34.

Gaffiot, F. (1934), *Dictionnaire latin-français*, Hachette, Paris.

Gay, H., O'Kill, B., Seed, K. and Whitcut, J. (1984, eds.), *Longman Dictionary of the English Language*, Longman, London.

Hajič, J., Panevová, J., Buráňová, E., Urešová, Z. and Bémová, A. (1999), *Annotations at Analytical Level: Instructions for annotators* [available online at *http://ufal.mff.cuni.cz/pdt/Corpora/PDT_1.0/References/aman_en.pdf*].

Korkiakangas, T. (2016a), *Morphosyntactic realignment and markedness change in Late Latin: Evidence from charter texts*, in «Pallas», 102, pp. 287-296.

Korkiakangas, T. (2016b), *Subject Case in the Latin of Tuscan Charters of the 8th and 9th Centuries*, Societas Scientiarum Fennica, Helsinki.

Korkiakangas, T. (2017), *Spelling variation in historical text corpora: The case of early medieval documentary Latin*, in «Digital Scholarship in the Humanities», 33, pp. 575-591.

Korkiakangas, T. (2018), *Spoken Latin behind written texts: Formulaicity and salience in medieval documentary texts*, in «Diachronica», 35, pp. 429-449.

Korkiakangas, T. (in press), *Late Latin Charter Treebank: Contents and annotation*, in «Corpora», 16, 2.

Korkiakangas, T. and Lassila, M. (2013), *Abbreviations, fragmentary words, formulaic language: Treebanking medieval charter material*, in Mambrini, F., Passarotti, M. and Sporleder, C. (2013, eds.), *Proceedings of the Third Workshop on Annotation of Corpora for Research in the Humanities*, Bulgarian Academy of Sciences, Sofia, pp. 61-72.

Korkiakangas, T. and Passarotti, M. (2011), *Challenges in annotating medieval Latin charters*, in «Journal of Language Technology and Computational Linguistics», 26, pp. 103-114.

Lewis, Ch.T. and Short, Ch. (1879), *A Latin Dictionary*, Clarendon Press, Oxford.

Longrée, D. and Poudat, C. (2010), *New ways of lemmatizing and tagging Classical and Post-Classical Latin: The* LATLEM *Project of the* LASLA, in Anreiter, P. and Kienpointner, M. (2010, eds.), *Proceedings of the 15th International Colloquium on Latin Linguistics*, Institut für Sprachwissenschaft der Universität Innsbruck, Innsbruck, pp. 683-694.

McGillivray, B. (2014), *Methods in Latin Computational Linguistics*, Brill, Leiden / Boston.

Murphy, M. (2010), *Meaning variation: polysemy, homonymy, and vagueness*, in Murphy, M. (2010, ed.), *Lexical Meaning*, Cambridge University Press, Cambridge, pp. 83-107.

Niermeyer, J.F., Van de Kieft, C. and Burgers, J.W.J. (2002, eds.), *Mediae Latinitatis Lexicon Minus* (revised edition), Brill, Leiden.

Philippart de Foy, C. (2012), *Lemmatiser un corpus de textes hagiographiques: enjeux et modalités pratiques*, in Biville, F., Lhommé, M.-K. and Vallat, D. (2012, *éds.*), *Latin vulgaire - Latin tardif IX. Actes du IX^e^ colloque international sur le latin vulgaire et tardif, Lyon, 2-6 septembre 2009*, Maison de l'Orient et de la Méditerranée 'Jean Pouilloux', Lyon, pp. 481-490.

Schiaparelli, L. (1929), *Codice diplomatico longobardo*. Vol. 1: *Fonti per la storia d'Italia 62*, Tipografia del Senato, Roma.

Schiaparelli, L. (1933a), *Codice diplomatico longobardo*. Vol. 2: *Fonti per la storia d'Italia 63*, Tipografia del Senato, Roma.

Schiaparelli, L. (1933b), *Note diplomatiche sulle carte longobarde*, in «Archivio storico italiano», 19, pp. 3-66.

Smith, J.Ch. (2011), *Change and continuity in form-function relationships*, in Maiden, M., Smith, J.Ch. and Ledgeway, A. (2011, eds.), *The Cambridge History of the Romance Languages*. Vol. 1: *Structures*, Cambridge University Press, Cambridge, pp. 268-317.

Sornicola, R. (2017), *La morfologia nominale: polimorfismo e polifunzionalismo nei sistemi di flessione*, in Sornicola, R., D'Argenio, E. and Greco, P. (2017, *a cura di*), *Sistemi, norme, scritture: la lingua delle più antiche carte cavensi*, Giannini, Napoli, pp. 85-134.

Väänänen, V. (1981, [1963[1]]), *Introduction au latin vulgaire*, Klincksieck, Paris.

Wright, R. (2000), *Latino e romanzo: Bonifazio e il Papa Gregorio II*, in Herman, J. and Marinetti, A. (2000, *a cura di*), *La preistoria dell'italiano: atti della Tavola Rotonda di Linguistica Storica (Università Ca' Foscari di Venezia, 11-13 giugno 1999)*, Niemeyer, Tübingen, pp. 219-229.

Treebanks

LLCT1 = https://zenodo.org/record/3633607#.XjU4lSNS9EY.

LLCT2 = https://zenodo.org/record/3633614#.XjU6zCN7lEY.

Timo Korkiakangas
Department of Languages
University of Helsinki
Unioninkatu 40
00014 Helsinki (Finland)
timo.korkiakangas@helsinki.fi

SSL

L.A.S.L.A. and Collatinus: A convergence in lexica

Philippe Verkerk, Yves Ouvrard,
Margherita Fantoli, Dominique Longrée

Abstract

The research group *L.A.S.L.A.* (Laboratoire d'Analyse Statistique des Langues Anciennes, University of Liège, Belgium) began in 1961 a project of lemmatization and morphosyntactic tagging of Latin texts. This project continues with new texts lemmatized each year (see *http://web.philo.ulg.ac.be/lasla/*). The resulting files, which contain approximatively 2,500,000 words, whose lemmatization and tagging have been verified by a philologist, have recently been made available to interested scholars. In the early 2000's, Collatinus was developed by Yves Ouvrard for teaching. Its goal was to generate a complete lexical aid, with a short translation and the morphological analyses of the forms, for any text that can be given to the students (see *https://outils.biblissima.fr/fr/collatinus/*). Although these two projects look very different, they met a few years ago in the conception of a new tool to speed up the lemmatization process of Latin texts at *L.A.S.L.A.* This tool is based on a concurrent lemmatization of each word by looking for the form in those already analyzed in the *L.A.S.L.A.* files and by Collatinus. This lemmatization is followed by a disambiguation process with a second-order hidden Markov model and the result is presented in a text-editor to be corrected by the philologist.

Keywords: lemmatization, morphosyntactic analysis, disambiguation, probabilistic tagger.

1. L.A.S.L.A.

The Laboratory for Statistic Analysis of Classical Languages (*L.A.S.L.A.* in the following) was founded in November 1961 at the University of Liège, by L. Delatte and E. Évrard. Its original aim is to lemmatize and analyze (tag) literary classical texts, both in Greek and in Latin, in order to produce indexes and to allow the study of classical languages with statistical and quantitative methods. This project, which is still on going, has already produced a large digitalized, lemmatized and annotated Latin corpus. This corpus covers the classical period, from Plautus to Ausonius, with some other Late-Latin texts. The *L.A.S.L.A.* Encoding Initiative interface allows the addition of new texts to the corpora. *L.A.S.L.A.* also released Textual

Received: January 2020
Accepted: May 2020

Data Analysis tools to access the information contained in its files (amongst which, for instance, the software Hyperbase; see *http://hyperbase.unice.fr/hyperbase/*). Through a specific agreement, access to these files is now free and open for every scholar who requests it.

1.1. *The structure of the files*

The *L.A.S.L.A.* Latin files contain fully lemmatized texts with a complete morphosyntactic analysis and some syntactic information. They have been systematically verified by a confirmed Latinist (either M.A. or Ph.D.). The annotation is not related to any specific grammar or to any specific linguistic description. In short, the available files are put in a text format where each line contains all the information related to a single token. As a reminiscence of the old punched cards, the fields have a fixed length, the blank character filling the empty spaces.

For each token of the text, the line begins with a unique alphanumeric code that identifies the text and a number that indicates the sentence count. All punctuation, which has been added by modern editors, is removed, except for the period that separates the sentences. The line then contains the lemma – as it appears in the dictionary of reference[1] – associated with an index if there are different homographs or to mark proper names or their derived adjectives. Then comes the form as it appears in the text, the reference – according to the *ars citandi* – and the complete morphologic analysis in an alphanumeric format[2]. For the verbs, an extra field, which remains empty for the other Parts-of-Speech (PoS in the following), gives some syntactic information: the verb of the main clause is identified and a subordinate code – depending on the subordination type – is affixed for the other verbs in the sentence.

The lemma always refers to an entry in the Forcellini's dictionary with a systematic disambiguator. For instance, POPVLVS_1 (i.e. *pŏpŭlus*, i, m.) is the people, while POPVLVS_2 (i.e. *pōpŭlus*, i, f.) is the poplar[3]. The PoS is also used to distinguish the homographs as AMICVS_1, the substantive, and AMICVS_2, the adjective. A problem arose for late Latin texts where an adjective can become a substantive. This is the case for SANCTVS,

[1] Cf. Forcellini (1864).

[2] As a matter of fact, two alphanumeric encodings co-exist, one in 5 characters – which is the original one – and the other with 9 – which is simpler. The matching can be done automatically.

[3] The two words are differentiated by vowel length and gender. POPVLVS_1 (*pŏpulus*), masculine, means "people" while POPVLVS_2 (*pōpulus*), feminine, means "poplar".

which is only an adjective in Classical Latin, but became a substantive later, especially in religious texts. To handle this situation an extra tag has been introduced: 'use as a substantive'.

During the tokenization process, the enclitics are separated from the rest of the form, but a special character is inserted in the line as a reminder that those two tokens correspond to a single word. Conversely, the encoding allows the treatment of verbal compound forms and also ellipsis. Crasis is treated in a way quite similar to enclitics: one word leads to two lemmata. Tmesis and compound words are also encoded in a special way.

The 9-character morphologic tag begins with a one letter PoS (A=noun, B=verb, C=adj, etc.), followed by a figure indicating the declension (for a noun), the conjugation (for a verb) or the class (for an adjective). Then come single digits indicating, if relevant, the case, number, degree, mood, tense, voice and person. For the same lemma, the figure indicating the declension can vary. For instance, *Vlixes* belongs, in principle, to the third declension. However, in accusative singular, the two forms *Vlixem* and *Vlixen* exist and are associated to different tags: A331 for the first one, as it is the normal Latin form, and A731 for the second form which is the Greek one. For the genitive, the two forms *Vlixi* and *Vlixei* are characteristic for the second declension, so the tag is now A241, although the lemma is still VLIXES. The gender is an extra piece of information but, due to the original decision made by the founders, it is not given for nouns and is not fully disambiguated. As a matter of fact, there are six possible genders according to the *L.A.S.L.A.* files[4].

1.2. *The* L.A.S.L.A. *Encoding Initiative interface*

The *L.A.S.L.A.* Encoding Initiative interface (see *http://cipl93.philo.ulg.ac.be/LaslaEncodingInitiative/*) is mainly a selection interface. A new text is first given to an operator who proceeds to some preprocessing[5]: tokenization, lemmatization and analysis. Until 2019, this was achieved with an analytic lemmatizer: starting from the ending, the lemmatizer broke the form down into morphemes in order to get all the possible roots and

[4] The three real genders and the three combinations that exist in the declensions (the f. + n. combination does not exist). We have plans (not yet fully implemented) to add the gender of the nouns and to disambiguate, when possible, the gender of adjectives, depending on the associated noun. Part of the task can be done automatically, but the result will have to be checked. Some words, as *canis* or *pereger*, are common (both masculine and feminine) and, even with the context, it may happen that the gender cannot be decided with certainty.

[5] See Denooz (1978) and Philippart de Foy (2014).

then recomposed in order to get all the possible analyses for all the possible lemmata. But the software supporting this lemmatizer was obsolete and *L.A.S.L.A.* decided to build a new lemmatizer based on form recognition. It means that each word of the text is compared to all the forms present in the *L.A.S.L.A.* forms dictionary. This forms dictionary includes all the possible forms for all the lemmata included in the *L.A.S.L.A.* lemmata dictionary. These possible forms are generated with a software based on morphologic rules, which adds all possible endings to each root corresponding to a lemma from the *L.A.S.L.A.* lemmata dictionary. The limitation here is that only lemmata already found in treated texts are included in the dictionary.

After this preprocessing, the text is presented as a list of tokens followed by all the known lemmatizations and analyses, one per line, in the alphanumerical order of the tags (corresponding to a given and fixed order of PoS, PoS-subcategories and morphosyntactic categories). Then, the philologist comes into play by selecting the 'correct one'. He/she is also invited to enter the syntactic information for the verbs, as it cannot be guessed by the computer. If a form is not in the dictionary or if the proper analysis is not given, the philologist has to add the correct analysis. The validation of the annotated text is possible only when the philologist has selected one analysis for each form of the text. At the end, the treated text returns to an operator who puts it in its final form.

Such a procedure ensures that the philologist has checked the lemmatization and the analysis of each token. As the computer does not select a priori a solution (even if there is only one possible lemmatization and analysis), the philologist has to read every line on the screen. But this process has its drawbacks, especially for technical texts with a specific vocabulary. As the dictionary has been built on Classical Latin literary texts, such as historical works, speeches, poetry, etc., a large amount of technical and scientific Latin words are missing from the *L.A.S.L.A.* dictionary and the philologist has to add them one by one[6]. Moreover, when the philologist inserts a new analysis into the lemmatization and tagging interface, there is no way to copy automatically this new analysis to the same form which could appear further in the text. As a matter of fact, to guarantee the coherence of its dictionary, *L.A.S.L.A.* does not update it automatically with the new forms.

[6] Indeed, the lexical specialization of technical and scientific texts, like didactic works and treatises on specific topics, is a feature which has been studied under many points of views, see for example De Meo (2005) and Fögen (2011).

1.3. *Access to the information*

There are several ways to access the information stored in *L.A.S.L.A.* files. The simplest approach is given by the interface Opera Latina[7] which allows for documentary search (indexes) but gives no statistics. A second possibility is to download the package Hyperbase-Latin which allows documentary and statistical exploration. This software has been developed in collaboration with Étienne Brunet of the laboratory Bases, Corpus, Langage (UMR 7032; CNRS-University of Nice)[8].

A more flexible approach is offered by the Hyperbase Web Edition interface[9]. One can choose between various databases or corpora. Beyond the usual documentary search (indexes), one can also ask for pattern detection – for instance all the sequences of two nouns. The Hyperbase Web Edition allows statistical searches such as z-score, factorial analysis or tree analysis. It is also possible to study the co-occurrences and even co-occurrences of pairs. As an extension of the Hyperbase Web Edition, HyperDeep, which is based on a Convolutional Neural Network, allows the identification of what is characteristic of a text or to find influences between authors.

For more specific purposes, *L.A.S.L.A.* files can be converted to XML and treated with TXM[10] or with data-mining tools[11].

2. *Collatinus*

Collatinus[12] was originally developed by Yves Ouvrard for teaching. It allows the generation of a complete lexical aid, with a short translation and the morphological analyses of forms, for any text which can be given to the students. As time went by, the lemmatizer has been augmented with other useful tools[13]. By simply clicking on a word, one can open a digital dictionary, e.g. Lewis and Short (1879) or Gaffiot (2016), to have the complete definition of the lemma. Another possibility is to scan a text to identify its

[7] Cf. *http://web.philo.ulg.ac.be/lasla/opera-latina/*. The list of the available texts is given at *http://web.philo.ulg.ac.be/lasla/textes-latins-traites/*.

[8] Cf. *http://web.philo.ulg.ac.be/lasla/hyperbase/*.

[9] Cf. *http://hyperbase.unice.fr/hyperbase/?edition=lasla*.

[10] Cf. *textometrie.ens-lyon.fr/spip.php?rubrique96*.

[11] Cf. *https://tal.lipn.univ-paris13.fr/sdmc/*.

[12] See Ouvrard and Verkerk (2014).

[13] For more details about these functionalities, see the article Ouvrard and Verkerk (in press), available as preprint at *https://hal.archives-ouvertes.fr/hal-02385036*.

metrical structure. A probabilistic tagger, based on a second order hidden Markov model (shorten as *HMM* in the following), allows the selection of the best lemmatization and analysis for each form by taking into account its context.

The lemmatization of a form is obtained by trying to split it as a root associated with a standard word-ending, which reproduces what the human reader does. The advantage of a program like Collatinus is that it is able to recognize forms not yet seen as soon as the root-word is known[14]. It is also easier to improve its base of knowledge: adding data for a new root-word allows the immediate recognition of ten or more (even a hundred, for verbs) forms[15]. Obviously, a program like Collatinus 'knows' a lot of forms that are not attested in the texts that have survived[16].

2.1. *Principle of operation*

When a student learns Latin, the first thing he/she has to understand is the way forms are constructed. Words are connected to an inflection paradigm. For each paradigm, one has to learn the list of word-endings and the rules to combine these endings with the roots that can be calculated, in some cases, or must be given. Collatinus works exactly in this way: one file provides the word-endings and the construction rules for each paradigm while another file connects the lemmata to the paradigms and provides also the roots which cannot be constructed. With this data, the construction of the inflected forms is immediate.

The lemmatization of a form requires the reverse process. For a given form, we have to split it in all the possible ways and to check that the first part coincides with a known root and the last one with a word-ending associated to the paradigm of the root[17]. The word-endings carry part of the information for the analysis, which is then stored in the file. Instead of an explicit analysis as e.g. 'nominative singular', we made a list of morphosyntactical analyses, which are possible in Latin and coded the analysis with a

[14] For any unknown form coming from an unknown root-word, it should be possible to guess a reasonable root-word in some simple cases.

[15] As it was the case before for the original *L.A.S.L.A.* lemmatizer.

[16] Note that, if the classical corpus is well established, it is not the case for medieval Latin.

[17] Going further, one can imagine to guess the lemma simply by subtracting the common word-endings. However, it would lead to surprising results. For instance, the form *merobibus* could be analyzed as an ablative plural of an hypothetical *merobis*. But such a method could give good results if several forms of the same lemma are found in a text.

simple number. As a matter of fact, the total number of these possible analyses amounts to 416. The number is converted into its human readable form when needed, i.e. for the display. Moreover, this encoding also allows the translation of the analysis into different languages[18].

2.1.1. *First difficulties*

One of the aims of Collatinus is to treat a Latin text as it is, without requiring some preprocessing steps like tokenization. A difficulty appears because of the enclitics *-que*, *-ne* and *-ve*. These words may be appended at the end of any form, and have to be separated before lemmatization. In most of the instances, the enclitics *-que* and *-ve* do not lead to ambiguous forms[19], which is not the case of the enclitic *-ne*. For instance, a form as *mentione* could be analyzed as the ablative singular of *mentio*, *onis*, as well as the nominative followed by the enclitic *-ne*. Enclitics, however, are not so frequent. We therefore assume that, if a form can be lemmatized as it is, then it is not necessary to search for the enclitics. In other words, the form *mentione* is now analyzed only as the ablative of *mentio*.

Collatinus also knows some contraction and assimilation rules. For instance, a double *i* appearing in the flexion of a word[20] is frequently written as a single long *i*. Some forms of the perfect can be contracted, the *-vi-* disappearing in, for instance, *amasse* (for *amavisse*). These forms are recognized by Collatinus, without the necessity of adding new word-endings. For the verbs constructed with a prefix, assimilation can change the spelling in some cases. It is the case, for instance, of *adfero*, *adtuli*, *adlatum* which often becomes *affero*, *attuli*, *allatum*[21]. The main assimilations of the prefix are known by Collatinus and built-in, so that it avoids the proliferation of forms for the same word.

2.1.2. *Distinction between* u *and* v

Very often, Latinists do not distinguish the letters *u* and *v*, and erase the *j* from the alphabet. But for scansion or counting syllables, it is clearly neces-

[18] For the moment, French, English and Spanish. But one can convert it to any other computer-oriented forms.

[19] A noticeable exception is *quo-que* that appears 7 times in the texts lemmatized by the *L.A.S.L.A.* (to be compared to the 2.290 occurrences of the lemma *quoque*).

[20] The first *i* ending the root, often short, and the second one at the beginning of the word-ending combine in a long *i*.

[21] Gaffiot (2016) gives the first forms, while Lewis and Short (1879) prefers the second ones.

sary to make a distinction. Thus, Collatinus keeps, in its lexicon and in the word-endings, the two consonants *v* and *j*, said to be Ramist consonants[22]. By the way, if one wants to use only *u* and *i*, it is easy to replace *v* by *u* and *j* by *i*. The proof, if needed, that preserving the distinction is the best choice is that the reverse process (restoring *v* and *j*) is almost impossible, and at least very difficult, except through a lemmatization method.

On the other hand, several Latin texts use only the *u* and *i*, and Collatinus knows this[23]. The solution to this problem is obtained through two steps. In a first step, all the *v* are replaced by *u* for the lemmatization. Then in a second step, the form is reconstructed from the root and the word-endings that eventually contain the *v* and *j*. As a result, a word as *uoluit* is analyzed as a form of perfect of either *volo* or *volvo*[24]. But if the text contains *voluit*, with a *v*, one can assume that it is not the perfect of *volvo*, otherwise it should have been written *volvit*, with two *v*'s. If the form of the text contains one (or more) *v*, the program eliminates any lemmatization that would lead to a reconstructed form with a different number of *v*'s.

Another class of *u* are not 'real' vowels, e.g. *suavis* or *sanguis*. It is also the case for the group *qu*, but in this group, the *u* is never a vowel. In the groups *sua* or *gui*, there are examples where the *u* is a vowel, for instance the possessive *sŭă* and the adjective *ām*bĭgŭīs[25]. It would have been shocking to write *svavis* or *sangvis* to stress that these words have only two syllables. Instead, we use the punctuated *-u* and write *sụāvĭs* and *sāngụĭs*[26].

2.1.3. *Word-endings and construction rules*

As already said, besides the lexicon which will be discussed later, Collatinus has another important file which gives the word-endings and the construction rules. For each paradigm, it gives the list of analyses and the

[22] Pierre de la Ramée (Petrus Ramus) is known in France to have introduced this distinction *u*/*v* and *i*/*j* in his *Gramere* (1562). But it seems that this idea appeared earlier in Spain (Antonio Nebrija, 1492) or in Italy (Giovanni Trissino, 1529). See Blanco and Bogacki (2014: 160 n. 24, 161).

[23] In the worst case, the editors write the capital *U* as *V*. It is not infrequent to find *Vnde* at the beginning of a sentence or to meet *Vlixes* in some texts.

[24] *Volvit* can also be a form of the present of *volvo*. The meaning of the sentence allows the reader to identify the correct form, but a computer does not understand the text. The case of *uoluit* can be a problem in prosody as it can counts for two or three syllables.

[25] The vowels are marked with a macron '¯' when they are long, as *ā* or *ī*, and with a breve '˘' when they are short, as *ĭ* or *ŭ*.

[26] Once again, if one does not want to use this strange character, it is easy to replace it by the standard *u*.

corresponding word-ending. A noun that follows a usual declension has 12 analyses and word-endings (some of them are identical), while an adjective has 108 possible analyses and word-endings. All the possible combinations of case, number, gender, degree, tense, mood and voice give 416 analyses which are just designated with a number. To avoid a very long enumeration of word-endings, we introduced a mechanism by which a paradigm 'inherits' the endings of its parent[27]. For instance, *miles* and *civis* have most of their endings in common, so we just have to indicate the differences.

Obviously, the word-ending is not the end of the story because one has to know the root to which this ending can be appended. For some declensions or conjugations, the roots can be calculated with just the lemma. For instance, for the first declension, it is sufficient to drop the last character of the lemma to have the root. In other cases, it must be given by the lexicon: one cannot guess the root *mīlĭt-* for the lemma *mīlĕs*. A more subtle example is the case of the first conjugation. In most cases, the roots for the perfect and the supine are obtained by adding *-āv-* and *-āt-* to the main root: the knowledge of the form *ămo* is sufficient to calculate the three roots *ăm-*, *ămāv- and ămāt-*, so it is not necessary to give them in the lexicon. But some verbs of the first conjugation do not follow this simple construction rule. To solve this problem, we have decided that if a root is given in the lexicon, it replaces the one that could be calculated. For instance, for the verb *sŏno*, we give the two roots *sŏnŭ-* and *sŏnĭt-* for the perfect and the supine.

2.1.4. *Ordering of the solutions*

For several forms, the result of the lemmatization is not unique[28]. Different words can lead to the same form, or a form corresponds to different analyses of the same word. Collatinus now gives the different solutions in an order that reflects the frequency of the use of the words. Up to version 10, the order of the solutions was alphabetical. As a result, the lemmatization of *suis*, for instance, gave the genitive of *sus*, *suis* as the first solution, although the ablative or the dative of *suus*, *a*, *um* are more likely.

[27] The construction rules are also transferred.

[28] There is a problem of vocabulary around the lemmatization: for the final user, the aim of a lemmatizer is to give *the* (unique) lemma associated to a given form in a given sentence. However, an operation that gives *all* the lemmata that can be associated with a form is also a lemmatization. We prefer to stick to this last sense and the full process with the association of a single lemma to a form is obtained with two steps: lemmatization and disambiguation.

Thanks to the statistics made from the lemmatized texts[29] of *L.A.S.L.A.*, we are now able to associate to each word of the lexicon a number of occurrences. Obviously, this number of occurrences is limited to the lemmatized corpus, but one can consider it as representative for the frequency of words. To go back to the previous example, *sus* appears 47 times in the texts of the *L.A.S.L.A.*, while *suus* appears 7,120 times. As Collatinus is not a form-lemmatizer[30], it does not know the number of occurrences for *suis* as dative plural of *suus* and for *suis* as ablative plural of the same *suus*. To order these two possible solutions, we make a strong assumption: the usage of the cases and number[31] (for nouns and adjectives; replaced by the mood for verbs) does not depend on the particular word. We still take into account the PoS[32] of the word. This evaluation does not reproduce exactly the observed frequencies, but remains a fair approximation. There are noticeable exceptions: for instance, *patres* is mainly a vocative plural, a case that is only very seldom used in other nouns/adjectives.

This ordering of the solutions is not sensitive to context. Its depends only on the form itself and its analyses. According to the statistics done on the lemmatized text of the *L.A.S.L.A.*, choosing the most frequent analysis gives the correct result in 80% of the cases. To reach a lower error rate, one can develop disambiguation methods based on the tagging of the words. These methods take into account, very crudely, the context of the word. They will be discussed later.

2.2. *Extension of the lexicon*

The lexicon of Collatinus contains the lemmata associated to a known paradigm, the different root-words that cannot be calculated and various pieces of information, such as the number of occurrences of this lemma in the texts lemmatized by the *L.A.S.L.A.* The translations of these lemmata are given in distinct files (one for each language) so that the material necessary to inflect or analyse the forms is independent from the translations. It also allows the addition of more languages for translations without having

[29] We did the statistical work a few years ago, and some new texts have been added to the corpus, which are not taken into account.

[30] We shall come back later on that example through the *L.A.S.L.A.* tagger.

[31] Unfortunately, the lemmatization by the *L.A.S.L.A.* does not give precisely the gender of the adjectives.

[32] Mainly: noun, adjective, verb and pronoun, as categorized by the *L.A.S.L.A.*

to duplicate or to change the basic information for the inflection. The files are just plain text-files, so that they can be edited and modified by the user to give better results.

Up to its version 10.2, the lexicon of Collatinus was set-up manually, the words being typed in when they were found in new texts given to the students. It contained slightly fewer than 11,000 entries, which allowed the lemmatization a significant portion of classical texts. However, we have decided to improve it by working on the dictionaries in a digital form. The two main dictionaries we have used are Lewis and Short (L&S), converted in XML by the Perseus Project[33], and Gaffiot, converted in TeX by a team lead by Gérard Gréco[34]. We have also used Georges[35] and Jeanneau[36] in their HTML forms. All these dictionaries are part of Collatinus. Some extra pieces of information were also used[37].

The first part of this work has been to collate all the lemmata together with the morphological information and the translation in each dictionary. The precise tagging of L&S and of Gaffiot, although very different, allows the compilation of very rich databases. The translations were probably the most difficult part of the job. Sub-entries, such as adjectives that derive from a noun that is the headword, were collected too. Orthographical variants, often indicated in an abbreviated form (e.g. *affĕro*, *better adf-)*, were expanded and added to the base. This has been done automatically but checked afterwards. The internal variants, (e.g. *rĕverto*, *rĕvorto)*, have been especially difficult to treat, although they are rather intuitive for the human reader. Obviously, one has to acknowledge the imperfection of the tagging[38]: some tags are missing or do not include all relevant information.

To deal with this lack of information, we combine the databases drawn from the various dictionaries, on the principle that, if a supine-form is missing in L&S, we can find it in Gaffiot (or vice-versa). This combination requires the alignment of the files, especially for homonyms, and the elimination of redun-

[33] Lewis and Short (1879), encoded in XML by Perseus (*http://www.perseus.tufts.edu/*).

[34] Gaffiot (2016), see *http://gerardgreco.free.fr/spip.php?article47*. Thanks to Gérard Gréco, we had access to the file before its publication.

[35] Georges (1913).

[36] Gérard Jeanneau, *http://www.prima-elementa.fr/Dico*. This Latin-French dictionary is still evolving. For this work, we have used a version of 2013.

[37] The data from Collatinus itself, a short version of Gaffiot, Lewis (1890), and the headwords of the *Pocket Oxford Latin Dictionary*, i.e. Morwood (2012).

[38] Here, we are considering the XML/HTML tags that identify the different entities. Later on, the word 'tag' will have a rather different meaning.

dant doublets. For instance, in L&S, *abscisus* has its own entry with a laconic definition «*P. a., v. abscido*» and is translated in a sub-entry of *abscido*. A supervised program allowed us to do this in a reasonable amount of time. Quantities can be sufficient to distinguish homonyms as *pŏpŭlus* vs *pōpŭlus*, but not always. Sometimes, we have to consider the PoS, as for instance in *a-spergo*, *ersi*, *ersum*, *3*, *v. a.* vs *aspergo*, *ĭnis*, *f.*, or the gender to recognize homonyms, for instance the noun *par*, *paris* which can be masculine or neuter. As a final option, the human reader can use the translations to align the entries.

The last step is to convert the collected information into a file which can be understood by Collatinus. The quantities given by the dictionaries are compared, and if they differ, we choose the form given by the 'majority'[39]. The quantities that can be determined by position are usually not indicated, but the program knows the rules[40] so that it was able to supply the missing quantities to Collatinus. Once again, a difficult step is the reconstruction of the roots: for the verb *a-spergo*, the program builds the form *āspērgo*[41] and the two roots, for the perfect and the supine, *āspērs*[42], while for the noun, it gives *āspērgŏ*[43] and *āspērgĭn*.

This treatment, mostly automated, yields to a lexicon of about 77,000 lemmata, associated with a paradigm and the necessary roots. But some 7,200 additional words were extracted from the dictionaries but not 'understood'. Some of them are useless for Collatinus: for instance, Gaffiot and the elementary Lewis have an entry for *aberam*, which is not a fundamental word. A Latinist should go through this file to determine which words may be useful to complete the lexicon. On the other hand, the process of expanding the variants of the headwords, which was necessary to align the entries of the dictionaries[44], leads to doublets. Most of the doublets caused by the assimilation of a prefix have been tracked down and suppressed. The Latinization of Greek names (e.g. *Ariadna*, *ae* for *Ariadne*, *es*) also caused

[39] In the comparison of quantities, we have to take into account that Georges (1913) and Lewis (1890) indicate only long vowels. The unmarked vowels can be either long by position or short.

[40] A diphthong is usually long (except for the *æ* of *præ* before a vowel, which becomes short). A vowel placed before two or more consonants is long too. A vowel before another vowel is short.

[41] The quantity of the final o is not relevant, because it is given by the word-endings.

[42] In these cases, the two roots are equal, but they usually differ. A difficult example is *ab-sorbĕo*, *bui*, *rarely psi*, *ptum* where we have two different roots for the perfectum, *ābsōrbŭ* and *ābsōrps*.

[43] The rule that says that the final *o* of the nominative is long when the previous vowel is long – see Quicherat (1885: 32), which can be downloaded from Gallica – seems not well followed. We prefer to mark it as common.

[44] For instance, Gaffiot (2016) has *adfero* as a headword, while Lewis and Short (1879) give *affero* with the variant *adf-*. Both are merged in Collatinus to give a single entry.

doublets. But a similarity of *a*/*e* or *us*/*os* is not sufficient to cause a doublet: for instance, *Agylla*, *ae* is an Etrurian city, while *Agylle*, *es* is a nymph. A final group of doublets comes from the singular or plural forms of some words which are chosen as headwords in the different dictionaries. A careful search for all of these doublets is still to be done.

Finally, to avoid long loading times, we split the lexicon into two parts. About one third of it corresponds to the 24,000 words that have been found in the texts lemmatized by *L.A.S.L.A.* It is loaded by default and allows the lemmatization of a large percentage of words in classical texts. The remaining two thirds, 53,000 words, are rarer words and are loaded only on demand. We planned to split the lexicon into more parts, each one specialized in a period of time or a range of semantically similar topics. We are considering this possibility for future versions as it requires that the program is able to load and purge different lexica while running[45].

2.3. *Perspective - Modularity of the data*

The 12th version of Collatinus (C12 here) is still under development. It focuses essentially on lexical and morphological data. Its aim is to handle larger and more precise data to lemmatize specialized corpora. For instance, when having to lemmatize a large medieval corpus, we confronted several difficulties:

- Numerous new words
- Evolution of semantics
- Evolution of graphic uses
- Evolution of paradigms

So, we found that the actual state of Collatinus' data often leads to wrongs results.

2.3.1. *Modules*

Our plan is to collect all the differences between the classical data and those which are required to lemmatize a non-classical corpus, for instance a medieval one. Using a special editor, a new set of data is created, containing all the differences between the classical state of the Latin language and the one in the corpus under study. These differences may appear at various levels: lexicon and translations, inflections, graphic usages, irregular forms.

[45] For the moment, Collatinus loads the data when booting.

This data is zipped into a package with the *.col extension. Once created, this module can be uploaded to the web-site of Collatinus. Then, other users can download it and install it in their C12.

Then, when lemmatizing a medieval text, the C12 user selects the medieval module. First, C12 reads classical data. Then, from this medieval module, new words are added. If a word already exists in the classical data, it is replaced by the medieval one. Often, the medieval word has few differences with the classical one: for instance, a new meaning. Sometimes, a word only needs to change its flexional paradigm, or one of its stems. But it may also be completely different. The same principle is applied for inflexions, irregular forms and graphic variants.

Orthographic variants: C12 adds a new data file, named *vargraph.la* which stores the orthographic particularities:

- Classical orthographical variants, e.g. *cu*/*quu* (*cum*/*quum*)
- Medieval orthographic variants are numerous, e.g.:
  - ligatures *q*;/*que*
  - phonetics *mpn*/*mn* (*dampnum*); *ß*/*ss*
  - tilde *ã* or *ā*/*an*, *am*

For medieval modules, the problem of the lexicon is very acute. Medieval corpora introduce many anthroponyms, toponyms, Latinization of local words: Celtic, Germanic, Spanish, etc. And these new words depends strongly on the considered corpus. For instance, the words derived from the vernacular languages will differ in Spain and in Germany. Thus, specific specialized lexica may be needed for each corpus[46].

A real difficulty is the survival of the anterior states of language. Classical authors could not know words to be created during the following centuries, but subsequent authors did know classical authors, sometimes very well. We need to be very careful when editing a classical word: classical senses may survive in medieval texts.

2.3.2. *The editor:* Ecce

Ecce (Ecce Collatinistarum Communitatis Editor) aims to create modules for C12. *Ecce*'s interface has four tabs: Lexical Modules, Lexicon,

[46] Another possibility would be to use an expandable personal lexicon, but it would remain 'private' and every scholar would have to develop their own lexicon. A third way could be to gather a huge data-base, but at some point a trade-off has to be made between the size of the base and the responsiveness of the program.

(ortho)Graphic variants, Irregulars. When launched, the first tab, Lexical Modules, is selected. On the left side, the user can choose the module to activate, deactivate, delete, generate or install. He can also choose other modules to extract data he will be able to add to the new module. Let us call them 'tank modules'. A very important tank is lem_ext, named 'extension'. When the new module and tank modules are selected, the user clicks the 'Activate' button. If this modular approach is adopted and widely used, the number of tank modules will grow, and building new modules will be easier and easier.

The Lexicon tab then appears. Latin text, and navigation buttons: beginning, backward, forward, previous failure, next failure, end. To feed the lexicon, the user clicks the 'next failure' button. *Ecce* goes on lemmatizing the text word after word, and stops when the lemmatization fails. The word is displayed, solutions, if any, are searched for in tank modules, so that you can check them, edit one of them, and add it. You can also, on the right side, edit a new lemma from scratch. If the lemma exists with another spelling, or another flexion, the two other tabs can be used. When the new data is validated, it is a good practice to go back to the beginning, and restart the lemmatization, to check if the edition is correct.

2.3.3. *Usages*

Collatinus is a lemmatizer, and its main usage is lemmatization. The modular organization of C12 allows a more precise lemmatization of non-classical or special corpora: author, place, topic. Just as Mario Nizzoli, in 1734, released a *Thesaurus Ciceronianus*, a Ciceronian C12 module could be created, uploaded to the web site of Biblissima and then downloaded by any other user who may be interested. It could be interesting to test it for teaching tasks:

- Provide a tiny module for a short Latin text;
- Ask students not to translate a text, but to develop the module which fits to this text, using *Ecce*.

3. L.A.S.L.A. - *Tagger*

As in every language, forms in Latin can be ambiguous. This ambiguity can be found at different levels. On one hand, in a declension, different cases can have the same form for the same word. A familiar example is the first

declension with the word-endings for the nominative and ablative which look the same but are different. On the other hand, some forms of different lemmata may coincide. For instance, *oris* is both a form of *ora*, *ae* and a form of *os*, *oris*. It can be useful to apply the usual techniques of disambiguation to propose the most probable analysis first. Obviously, one also has perfect homographs, as the two *populus* or the two *levis*, that share the same inflected forms and are completely undistinguishable.

3.1. *Statistics on lemmatized texts*

Methods based on 'hidden Markov models', commonly known as probabilistic taggers, are widely used for disambiguation of the modern languages[47]. They assign to each form a tag that reflects its morphosyntactic nature and sometimes its syntactic function. The PoS is often used as a tag, sometimes complemented with some other pieces of information. The method relies on the hypothesis that the sequences of tags are characteristic of the language and do not depend on the text, whatever the subject is and whoever the author. Knowing the frequencies of the pairs (form, tag) and the frequencies of the sequences of three tags (second order Markov process), one can compute the probabilities associated with each of the possible sequences of tags for the sentence. Then one assumes that the most probable sequence is the correct one, or at least the more likely one[48]. Very high accuracies are obtained with modern languages, where the order of the words in the sentence is rather fixed. It is not demonstrated that the same fidelity can be reached with Latin, where the order of the words is free, or at least much freer than in modern languages.

On the other hand, in the last decade, new methods appear which are based on Artificial Intelligence (AI) and, often, Neural Networks. They give better results than *HMM* with modern languages. However, the evaluation of the error rates is sometimes questionable, especially for Latin, as the 'tasks' for lemmatization and PoS tagging are separated. If an AI program analyzes the form *cum* as the accusative of a substantive *cum* (2nd declension neuter, obviously), it could be counted as a correct lemmatization[49]. Anyhow, even if the error rate of AI methods is lower, it is still far from the aim of the phi-

[47] See for instance Rabiner (1989).

[48] For a more detailed description of a tagger, see Schmid (1994), available at *http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/*.

[49] Except that *L.A.S.L.A.* knows several lemmata *cum* that have an index 1, 2, etc.

logist who wants to reach a golden standard result with no errors. Moreover, the AI methods require more computational power than *HMM* and behave as a magical black box. What we need is a practical and robust tool giving understandable results. *HMM* relies on a simple hypothesis and then one needs only Bayesian probabilistic calculations. The principle of *HMM* is easy to understand and to follow.

To begin, one has to choose the tag-set and to do some statistics on a training corpus[50]. A trade-off has to be made for the tag-set. If the tag-set is too small, its disambiguation capabilities will be restricted: for instance, if we just consider the PoS, we will not be able to distinguish the two *oris*, which are both nouns. On the other hand, if the tag-set is too large, the statistics on a finite corpus will be poor. As a training corpus, we used texts lemmatized and analyzed by *L.A.S.L.A.*[51]. The files we used count slightly fewer than two millions words, each form being associated with a lemma and a code that gives the full analysis[52]. This code cannot be used as a tag, because it would lead to an excessively large tag-set with more than 3,000 different tags. We cut from these codes some redundant information: for instance, for verbs, the type of conjugation is associated to the lemma and the different persons have different word-endings. We choose to restrict the tag to the PoS associated with the mood for verbs and with the case and number for the declined forms[53]. For each triplet (form, lemma, tag), we counted the number of occurrences in the corpus. We obtained a file with about 150,000 entries. And we did the same for the sequences of three tags, obtaining a file with 235,000 entries. These numbers are the primary information sources for the implementation of a probabilistic tagger.

3.2. *Double lemmatization*

With the statistical data extracted from the texts lemmatized by *L.A.S.L.A.*, we have developed a lemmatizer-tagger. The first version of the program began with a sequential lemmatization. It first looked if the form

[50] It is not a training corpus in the sense used today in Neural Networks and AI, even if Schmid (1994) called it training. It is a fully annotated corpus on which statistics are performed in a perfectly mastered way.

[51] We thank Gérald Purnelle for his help in the preparation of these texts, the list of which can be found at: *http://web.philo.ulg.ac.be/lasla/textes-latins-traites*.

[52] The gender is absent in the corpus we have treated.

[53] The number is needed only to distinguish some forms, mainly in the fourth declension and could be omitted. A lot of tests should be done to optimize the tagset, which have not yet been done.

was found in the file containing all the forms of the texts lemmatized at *L.A.S.L.A.* (form lemmatizer). If a word was not found in the file, the code sent a request to Collatinus which was supposed to run in the background on the same computer. Collatinus answered with the possible lemmatizations of this form. If Collatinus was not able to answer (either because it was not running or because it did not recognize the form), then the program asked the philologist who was supposed to supervise the process and waited for an answer.

However, since it has been found that this sequential and conditional lemmatization induces errors, we turned to parallel lemmatization[54]: the lemmatization is always done both by Collatinus and by a form-lemmatizer based on the *L.A.S.L.A.* data. The cause of the errors in sequential lemmatization was the fact that as soon as one solution was given by the form-lemmatizer, the program assumed that all the solutions were given. But consider, for instance, nouns where dative plural and ablative plural have the same form. It occurs frequently that for some lemmata only one of these two cases has been found in the *L.A.S.L.A.* texts. As a consequence, the program assumed that only one tag could be associated with this form, reducing erroneously the tag-sequences to be tried. Then the error propagates due to the mechanism of the probabilistic tagger, forcing the philologist to correct several analyses in the sentence.

The double lemmatization requires extra work to match, if possible, the lemmata used by *L.A.S.L.A.* with those of Collatinus and to remove the duplicates. The correspondence between the two lexica is rather delicate. Just to give a few examples, *L.A.S.L.A.* distinguishes the two *et*, conjunction or adverb, while Collatinus has a single lemma *et*, with two possible PoS. On the other hand, Collatinus considers (up to now[55]) that *poplus* is a lemma, while *L.A.S.L.A.* considers it as a contracted form of POPVLVS_1[56]. The correspondence has been established by asking Collatinus to lemmatize the

[54] Independently, Patrick Burns developed concurrent lemmatization (see elsewhere in this volume).

[55] In the last version of Collatinus, we have introduced the possibility of giving several forms for a lemma, but we have not yet reviewed the whole lexicon to group those forms.

[56] As a matter of fact, the lemmata in the lexicon of *L.A.S.L.A.* are given in uppercase, with a disambiguation index if necessary. By convention, proper names and the associated adjectives have always an index, N and A (sometimes O, if there are homonyms as *Pallas*, *adis*, f. and *Pallas*, *antis*, m.). Otherwise, the index is present only when there are homonyms and is an integer (1, 2, etc.). In Collatinus' lexicon, the lemmata are written as usual: in lowercase, with an index if there are homonyms (for historical reasons, the index 1 is generally omitted – which is probably not a good idea) and with a capitalized first letter for proper nouns and adjectives.

list of forms found in the *L.A.S.L.A.* files (as mentioned above, the form is associated with a lemma and a code giving the PoS and the analysis). The PoS and the analysis given by Collatinus were compared with the *L.A.S.L.A.* code. In the best case, the match is unique and perfect, and then the two lemmata are linked. Otherwise, a list of suitors is established and an algorithm tries to sort it out. At the end, a manual check has to be done[57].

As mentioned above, Collatinus does not split the enclitics *-que* or *-ne* if the word is recognized as a whole. So this possibility has been added in the editor of the annotated text. On the other hand, Collatinus does not search for compound verbal forms, so *amata est* will remain a participle followed by a verb, just as *fortis est* is an adjective followed by a verb. However, in the double lemmatization, if the compound form has been seen in the *L.A.S.L.A.* corpus (which is the case for *amata* <*est*>) then the program will offer this solution as the preferred one. This particularity may lead to apparent inconsistencies as, for instance, *est amatus* will be recognized as a compound verbal form while *amatus est* will not. But the philologist will have the ability to add any compound forms.

3.3. *Disambiguation*

The results are sorted by frequency, and a first attempt for the lemmatization of the text is obtained by putting together the most frequent individual lemmatizations. This first attempt considers the forms as isolated, independent of their neighbours, and its error rate is expected to be about 20%[58]. Then, the tagger enters play to take into account the context with a simple statistical model. We have made very few trials: the obtained accuracy was about 88% (exact result, i.e. correct lemma and analysis) and the lemma is the correct one in 96% of the cases. As a last step, the philologist can check all the lemmatizations and, if needed, correct them.

As already mentioned, we are not interested in having the lowest error rate for the tagger itself. The only aim is to facilitate the philologist's work with a convenient tool. We did not sacrifice part of the annotated corpus to keep a 'test corpus', so the evaluation of the tagger has to be done on excerpts

[57] As one has to deal with a few thousand lemmata, some errors remain in the look-up table. Some correspondences are also missing.

[58] This figure is evaluated on the training corpus. If we consider the most frequent lemmatization of each form and sum the corresponding numbers of occurrences, we obtain about 80% of the total number of lemmatized forms.

of this same corpus. Some will argue that it is cheating, but laws about entropy show that, when the corpus is large enough, the computer cannot remember all the sequences it has seen and the results will not change significantly. More interesting is the evaluation of the number of changes that the philologist did between the first attempt by the tagger and the final file when facing a completely new text. For example: an extract of Ausonius' *Mosella* with a total of 1,826 tokens[59]. Considering only the lemma and its index, we have observed 218 modifications of which 31 were due to changes in the text: the philologist erased a verse and corrected some OCR errors (e.g. *lam* corrected to *Iam*, *amatam* for *afflatam*). As the lemmata given by Collatinus are in lowercase, a normalization (to the uppercase lemmata used in *L.A.S.L.A.*'s corpus) is needed when the lemma is new to *L.A.S.L.A.* Such a normalization is not related to an error of the tagger[60] and the corresponding cases are excluded from the analysis. In the end, the mistakes of the tagger were 125, an error rate of about 7%. This sample is too small to analyze it statistically, but it turns out that a significant part of the mistakes are due to the ambiguity between participles and adjectives (in both directions, for instance, *compositus* vs *compono*, or *fulgo* vs *fulgens*) and sometimes between noun and adjective (for instance, *Alpinus*). Some errors are due to the mishandling of the capital at the beginning of a verse and could be corrected. More difficult is the case of the enclitic: we have chosen that if the form exists as a whole, we do not try to strip off the enclitic *que*, for instance in *quaque* which, sometimes, has to be split in *qua-que*. Another difficulty comes from *L.A.S.L.A.*'s fine-grained lemmatization: a simple form as *ut* is connected to four lemmata and *quo* to five (each lemma is associated with one PoS). A second analysis on Prudentius' *Psychomania* gives similar results on a sample of 6,133 tokens, and most of the errors are due to the uncertainty between participles and adjectives.

With a probabilistic tagger, it is interesting to note that, although the 'context' is described by the sequences of three tags, the choice of the best tags is done only at the end of the sentence or of the text. In principle, all the possible sequences of tags are considered, but many of them are skipped[61]. In any case, the choice of a tag can influence the analysis of another word further than two words apart. Conversely, it is important to know how far a 'wrong'

[59] This text is part of the work of Marc Vandersmissen for a research project F.R.S.-FNRS-PDR FNRS-2019: Motifs textuels ovidiens et littérature latine tardo-antique.

[60] We could have done the transformation a priori, but we wanted to single out these new lemmata. It allows the philologist to preserve the coherence of the lexicon.

[61] For details about the pruning method, see Schmid (1994).

analysis would spread its effect. An examination of the list of words shows that slightly less than 40% of the forms are associated with a unique analysis (thus a single tag). Thus, the probability of finding two such forms consecutively is 15%, which means that such a pair should be found, on average, every 6 or 7 words. Such a pair splits the text because these unique tags are present in all the tag-sequences, forming fixed points. The fact that we use a second order Markov model implies that the tags that come after a fixed point do not depend on the tags before. Therefore, if the tagger gives the wrong tag to a word, this error will affect some of the following words, but not many. Roughly speaking, it can affect seven words, on average. Obviously, it may happen that a longer series of words can be found between the fixing pairs.

One can imagine a 'multiplex disambiguation' with another method, which would allow for cross-checking the results. A huge benefit[62] can be achieved if the methods differ sufficiently, even if they are trained on the same corpus. Neural networks and AI are presently very promising in this direction. However, their outputs should be cleaned from the absurdities they can contain. For instance, it has been seen[63] that the output of a neural network program contains 'Cum ; cvm ; NOM2 ; Case=Acc|Numb=Sing': the form *cum* is analyzed as the accusative singular of a noun (lemma) *cvm* following the second declension. Clearly, some constraints have to be added to the program. One of the problems with AI methods (in general, this is not specific to this process) is that nobody knows why the program chose one solution instead of another one. This is not the case with *HMM* where the reason for the choice is always that a probability is larger than another one. By looking closer at these probabilities, it should be possible to associate a 'confidence level' to any result. If the larger probability differs from the second one by a small amount, then the confidence level is poor and the philologist should check the result twice. But this remains to be done, and it raises fundamental questions. For instance, what counts as a small difference in probabilities? How can the program, which does not understand what it is reading, know where the difficulties are?

From a more theoretical point of view, it would be interesting to study the sequences of tags to search for correlations. If the order of the words were completely free, one would expect no correlation at all and the tagger would

[62] However, for the philologist who wants zero error, it will not be sufficient. A careful and tedious check will always necessary.

[63] We shall not mention where.

give the same result as a frequency-based lemmatizer. The correlations and the efficiency of the tagger are linked, and the study of the former will give information on the limits in the accuracy. As for the previous point, this work remains to be done. And both points may well be correlated.

3.4. *Comparison*

The content of this section is mainly subjective and speculative. As a matter of fact, nobody will ever lemmatize the same text with each of the two proposed tools. It would mean to do twice the job with no benefit.

The traditional procedure for preparing *L.A.S.L.A.* files is semi-automatic: the lemmatizer proposes to the philologist all the analyses known by the *L.A.S.L.A.* dictionary for each of the forms in the text. The philologist selects the correct analysis, or inserts manually the correct analysis, if needed. The analyses are proposed in an order depending only on the morphosyntactic code, and not on their frequency or on their likeliness in that context.

On the contrary, the tagger proposes the most probable analysis, and therefore the role of the philologist is essentially to correct the results of the analysis proposed by the tagger. This accelerates the work, but also changes the kind of human mistakes that occur. On the one hand, the traditional *L.A.S.L.A.* procedure induces human mistakes caused by the similarity of the possible morphosyntactical analyses, represented by similar alphanumerical codes. The philologist may mistake an accusative for a nominative, or an ablative for a dative, or pick the wrong mood or tense for a verb. It is highly unlikely that, in case of homographic forms, like for instance *salis* (2nd person of the present indicative of *salio*, or genitive from *sal*), the user would select the verbal analysis instead of the nominal or vice versa. On the other hand, the tagger may be lead to such an erroneous choice, but the mistake shall remain unseen by the philologist. Indeed, since the philologists expects, for instance, a genitive, he may think that the form is unambiguous, because the possible analysis as the indicative of the verb *salio* may not occur to him in that context. Therefore, attention may lapse, and the tagger's mistake may be left unseen. With the traditional method, the user would hardly mistake the analysis of the verbal form with the one of a substantive. When using the tagger, on the contrary, the philologist is more conscious of the necessity of checking the proposed solution for clearly potentially ambiguous forms, such as datives/ablatives, and will thus probably pay high attention to the correction. At the moment it is not possible to verify which of the methods causes

more human mistakes, therefore it is not possible to draw any conclusion on this topic. The two methods are synthetically compared in Table 1:

L.A.S.L.A. *Encoding Initiative*	*Collatinus*-L.A.S.L.A. *tagger*
PREPARATION OF THE TEXT	
The text is prepared by an operator from *L.A.S.L.A.*	The text is loaded directly in the program, with a minimal standardization in the splitting of lines/paragraphs/chapters/etc.
PROS: Initial control of the edition, of the splitting, etc.	PROS: The philologist can start to work immediately. He/she has the possibility to correct/change the references and the text during the lemmatization.
CONS: Possible delays, independent of the will of the philologist.	CONS: Possible use of texts (for instance, available on internet) without any indication of the reference to the edition.
Comment: The tagger offers more flexibility, but requires more care and knowledge about the mechanisms of reference and the choice of the edition.	
CHOICE OF THE ANALYSES	
Proposition of all the known analyses, without any priority.	Proposition by default of the 'best' solution, together with all the other possible analyses.
PROS: The philologist has to read carefully all the given analyses to select one of them.	PROS: Fast processing and several cases are solved automatically.
CONS: Constant concentration (even for the simple cases). Slower treatment.	CONS: The default choice may be wrong and still escape the philologist's attention.
Comment: An evaluation of the error rates achieved with the two methods has to be done. It is a difficult task from a methodological point of view because it is not the philologist who is evaluated, nor the complexity of the considered text.	
DICTIONARY	
The dictionary is based on the Forcellini. The addition of new lemmata is controlled by the PI at *L.A.S.L.A.*	The dictionary is based on Gaffiot and Lewis & Short. A personal lexicon is added.
PROS: Internal coherence for the whole corpus of *L.A.S.L.A.* and also in the propositions given in the program.	PROS: More extended lexical base. New entries can be added simply. Distinction between lemmata known by *L.A.S.L.A.* (in uppercase) and those from Collatinus (in lowercase).
CONS: Frustration of the manual insertion of new lemmata/analyses. Risk of error in the repetition of this task.	CONS: Risk of incoherence with the *L.A.S.L.A.*'s corpus. Possibilities of unseen doublets or errors in the indices.
Comment: Strong advantage in the speed of the tagger. If the personal dictionaries were checked and inserted in the *L.A.S.L.A.* dictionary, it would increase its size rapidly.	

L.A.S.L.A. *Encoding Initiative*	*Collatinus*-L.A.S.L.A. *tagger*
Final treatment	
Usually, the treated text is checked (often by another philologist). Correction of the printed index and insertion of them by an operator. Production of the final file, by an operator, at the end of the process (for instance, several books).	The generation and the correction of the index are left to the philologist. The output file is immediately in the standard APN format which makes it usable at once.
Pros: Rigorous verification, in part on printed material.	Pros: The file can be studied as soon as it is completed, without having to wait for the completion of the entire work (if formed of several books).
Cons: Possible delays in the processing (in part independent of the philologist's will).	Cons: Risk of a less careful verification.
Comment: Working with the tagger appears to be a more personal work, with more responsibilities but more independence and flexibility.	
Conclusion: For a work to be completed in a finite amount of time (e.g. for a PhD thesis), the speed of the tagger is a key element. The philologist at work has a complete control of all the steps, but also (as a consequence) a larger responsibility. On a longer time scale, the traditional method is safer for the coherence of the *L.A.S.L.A.* corpus. However, nothing impedes an extra checking of the output of the tagger (by a second philologist) to ensure its quality. The coupling of the two methods could lead to a significant increase of the *L.A.S.L.A.* corpus and dictionary.	

Table 1. *Summary of the differences between the two NLP tools.*

4. *Conclusion*

In this article, we have presented part of the work going on at the *L.A.S.L.A.* and in the Collatinus' development group. We have also put some emphasis on their collaboration and compared the two approaches for the lemmatization and analysis of new Latin texts. We underline the pros and cons of each of them. A kind of trade-off has to be found between speed and precision.

However, the required precision or the tolerable error rate may depend on the envisioned application and remain an open question. Obviously, a perfect lemmatization, with no error at all, is desirable, but probably not needed. Most of the applications are of statistical nature, which means that they contain an intrinsic degree of uncertainty which can often be determined with error-bars, but seldom given or understood. In this context,

what is (or would be) the consequences of a few remaining errors? It is difficult to evaluate, but even more difficult to measure. Due to the lack of realistic objectives (with upper limits on the acceptable error rate, for instance), we stick to perfection.

Acknowledgements

We would like to warmly thank Bret Mulligan (Haverford College) for carefully reading our text and for his very pertinent remarks.

References

Blanco, X. and Bogacki, K. (2014), *Introduction à l'histoire de la langue française*, Bellaterra, Barcelona.

Denooz, J. (1978), *L'ordinateur et le latin. Techniques et méthodes*, in «Revue de l'organisation internationale pour l'étude des langues anciennes par ordinateur», 4, pp. 1-36.

De Meo, C. (2005), *Le lingue tecniche del latino*, Pàtron, Bologna.

Forcellini, E. (1864, [1771[1]]), *Totius Latinitatis Lexicon* [ed. by F. Corradini and G. Perin], Tipografia del Seminario, Padova.

Fögen, Th. (2011), *Latin as a technical and scientific language*, in Clackson, J. (2011, ed.), *A Companion to the Latin language*, Wiley / Blackwell, Oxford, pp. 445-463.

Gaffiot, F. (2016). *Dictionnaire latin-français par Félix Gaffiot, revu et corrigé sous la direction de Gérard Gréco* [available online at *http://gerardgreco.free.fr/spip.php?article47&lang=fr*].

Georges, K.E. (1913), *Ausführliches lateinisch-deutsches Handwörterbuch*, Hahnsche Buchhandlung, Hannover.

Lewis, Ch. T and Short, Ch. (1879), *A Latin Dictionary Founded on Andrew's Edition of Freund's Latin Dictionary*, Clarendon Press, Oxford.

Lewis, Ch. (1890), *An Elementary Latin Dictionary*, American book company, New York / Cincinnati / Chicago.

Morwood, J. (2012), *Pocket Oxford Latin Dictionary*, Oxford University Press, Oxford.

Ouvrard, Y. and Verkerk, Ph. (2014), *Collatinus, un outil polymorphe pour l'étude du latin*, in «Archivum Latinitatis Medii Aevi», 72, pp. 305-311.

Ouvrard, Y. and Verkerk, Ph. (in press), *Collatinus & Eulexis. Latin & Greek Dictionaries in the Digital Ages*, in «Classics@».

Philippart de Foy, C. (2014), *Nouveau manuel de lemmatization du latin* [available online at *http://hdl.handle.net/2268/162433*].

Quicherat, L. (1885, [1846][1]), *Nouvelle prosodie latine*, L. Hachette, Paris.

Rabiner, L.R. (1989), *A tutorial on hidden Markov models and selected applications in speech recognition*, in «Proceedings of the IEEE», 77, 2, pp. 257-289.

Schmid, H. (1994), *Probabilistic part-of-speech tagging using decision trees*, in *Proceedings of the International Conference on New Methods in Language Processing*, UMIST, Manchester, pp. 44-49.

Philippe Verkerk
Laboratoire de Physique des Lasers, Atomes et Molécules
Université de Lille
Bâtiment P5
F59655 Villeneuve d'Ascq Cedex (France)
Philippe.Verkerk@univ-lille.fr

Yves Ouvrard
Retired professor of 'Éducation Nationale'
Yves.Ouvrard@collatinus.org

Margherita Fantoli
Laboratoire d'Analyse Statistique des Langues Anciennes
Université de Liège
Place du 20 Août, 7
B-4000 Liège (Belgique)
mfantoli@uliege.be

Dominique Longrée
Laboratoire d'Analyse Statistique des Langues Anciennes
Université de Liège
Place du 20 Août, 7
B-4000 Liège (Belgique)
dominique.longree@uliege.be

SSL

The Frankfurt Latin Lexicon: From morphological expansion and word embeddings to SemioGraphs

Alexander Mehler, Bernhard Jussen, Tim Geelhaar, Alexander Henlein, Giuseppe Abrami, Daniel Baumartz, Tolga Uslu, Wahed Hemati

Abstract

In this article we present the Frankfurt Latin Lexicon (*FLL*), a lexical resource for Medieval Latin that is used both for the lemmatization of Latin texts and for the post-editing of lemmatizations. We describe recent advances in the development of lemmatizers and test them against the Capitularies corpus (comprising Frankish royal edicts, mid-6th to mid-9th century), a corpus created as a reference for processing Medieval Latin. We also consider the post-correction of lemmatizations using a limited crowdsourcing process aimed at continuous review and updating of the *FLL*. Starting from the texts resulting from this lemmatization process, we describe the extension of the *FLL* by means of word embeddings, whose interactive traversing by means of SemioGraphs completes the digitally enhanced hermeneutic circle. In this way, the article argues for a more comprehensive understanding of lemmatization, encompassing classical machine learning as well as intellectual post-corrections and, in particular, human computation in the form of interpretation processes based on graph representations of the underlying lexical resources.

Keywords: lemmatization, crowdsourcing, post-correction, stratified embeddings, SemioGraph.

1. *Introduction*

Regarding lexical resources for Natural Language Processing (NLP) of historical languages such as Latin, three paradigms can be roughly distinguished: (i) morphologically enriched lexica or dictionaries such as the Frankfurt Latin Lexicon (*FLL*) to be presented here, which use, for example, rules of morphological expansion to generate inflected forms from lemmas collected from web and other resources, (ii) wordnets such as the famous WordNet (Miller, 1995), which as a terminological ontology (Sowa, 2000) distinguishes (wordforms as search terms of) lemmata from synsets and their sense relations, and (iii) word embeddings (Komninos and

Received: January 2020
Accepted: April 2020

Manandhar, 2016; Levy and Goldberg, 2014; Ling *et al.*, 2015; Mikolov *et al.*, 2013) which address the statistical modeling of syntagmatic (contiguity) and paradigmatic (similarity) associations of lexical units (Halliday and Hasan, 1976; Jakobson, 1971; Miller and Charles, 1991; Raible, 1981)[1]. Ideally, for a historical language such as Latin, there exists an integrated system of resources of these kinds in a sufficiently deep state of development: by using such a resource, a professional user or NLP system is extensively informed about the lexical units of the target language on different levels of lexical resolution (including wordforms, lemmata, superlemmata, lexeme groups, etc.), about their morpho-syntactic and semantic features as well as about their various sense relations and unsystematic associations. Regarding the example of lemmatization, such a resource would support both the initial automatic lemmatization and its intellectual post-correction, which in turn would be the starting point for the post-correction or further development of this resource, thereby closing the digitally enhanced hermeneutic circle. However, the example of the Latin WordNet (Minozzi, 2017) already shows that the components of such an integrated resource are still out of reach for this historical language (as explained and analyzed in Franzini *et al.*, 2019). The same applies to input-intensive word embeddings, which require large amounts of text data, but which are not yet freely available for Latin (see, however, *UDify*, Kondratyuk and Straka, 2019, as an example of an approach that seems to circumvent this limitation – cf. Section 4). Last but not least, voluntarily created lexical information systems such as Wiktionary, which aim to integrate wordnet-related information with dictionary information, suffer not only from a lack of scope, but also from multiple sources of information biases (cf. Mehler *et al.*, 2017). Therefore, it remains a challenge to provide not only one of the three types of resources (dictionary, wordnet, embeddings) for Latin in sufficient quantity, but even more to do so for at least two of these types – in an integrated manner. This article wants to take a step in this direction. That is, we present the *FLL* as a kind of Latin dictionary that distinguishes lexical units at the level of word forms, syntactic words[2], lemmata, and superlemmata, provides rich grammatical in-

[1] In the case of modern languages, a fourth paradigm would be given by knowledge graphs derived, for example, from Wikidata or Wikipedia.

[2] Syntactic words are signs in the sense of structuralism (Saussure, 1916): they include an expression plane (called 'wordform') and a content plane. The content plane of syntactic words is usually represented by an attribute value-structure that collects grammatical features such as *case* and *numerus* in the case of nouns or *tempus* and *genus verbi* in the case of verbs. Thus, the same wordform may be

formation for syntactic words, serves as a resource for the post-correction of automatic lemmatization, provides a word-for-word monitoring of the lemmatization status of each text, which is particularly easy to read for non-IT philologists, and supports the computation of word embeddings at various levels of lexical resolution. We also show how these embeddings can be presented as interactive graphs to encourage the correction and further development of the underlying resources.

The article is organized as follows: Section 2 outlines the structure of the *FLL* and quantifies the extent of its overgeneration or, conversely, its lack of coverage. Section 3 deals with the post-processing of the lemmatization of Latin texts with the help of the *FLL*, while Section 4 compares the current progress of lemmatizers for Latin. Subsequently, Section 5 deals with the derivation of genre-sensitive word embeddings for Latin and their visualization by means of interactive SemioGraphs. These visualizations are then used in Section 6 to conduct case studies in computational historical semantics, which ultimately combine lemmatization and the evaluation of word embedding graphs with the underlying *FLL*. In this way, we will speak of a 'digitally enhanced hermeneutic circle' implemented through the NLP pipeline for Latin, as presented in this article. Finally, Section 7 draws conclusions and gives an outlook on future work.

2. *From superlemmata to syntactic words*

The Frankfurt Latin Lexicon[3] (*FLL*) is a morphological lexicon, currently for Medieval Latin, that is Latin between 400 and 1500 CE. Its main purpose is to support the automatic lemmatization of Latin texts with the Text-technology Lab Latin Tagger (*TTLab* Tagger) (Gleim *et al.*, 2019; cf. Stoeckel, 2020), which is available through the TextImager[4] (Hemati *et al.*, 2016), the eHumanities Desktop[5] (Gleim *et al.*, 2012) and GitHub[6]. It was created starting in 2009 (cf. Mehler *et al.*, 2011; see also Jussen *et al.*, 2007)

mapped to different syntactic words (as, for example, *house* in *Your house*$_1$ *is next to her house*$_2$ in which the tokens *house*$_1$ and *house*$_2$ manifest the same wordform but two different syntactic words distinguished by case).

[3] Cf. *https://www.comphistsem.org/70.html/*.

[4] Cf. *https://textimager.hucompute.org/*.

[5] Cf. *https://hudesktop.hucompute.org/*.

[6] Cf. *https://github.com/texttechnologylab*.

by extracting and collecting lemmata from various web-based resources. This includes[7] the *AGFL* Grammar Work Lab[8] (Koster and Verbruggen, 2002), the Latin morphological analyzer *LemLat* (Passarotti, 2004), the Perseus Digital Library (Crane, 1996), William Whitaker's Words[9], the Index Thomisticus[10] (Passarotti and Dell'Orletta, 2010), Ramminger's Neulateinische Wortliste[11], the Latin Wiktionary[12], Latin training data of the Tree Tagger (Schmid, 1994), the so-called Najock Thesaurus[13], and other resources from cooperating projects like Nomen et Gens that provide several thousands of Latin personal names[14]. Since then, the *FLL* has grown continuously through the lemmatization of Latin texts[15].

The entries of the *FLL* are structured according to a four-level model consisting of wordforms, syntactic words (mapping wordforms onto vectors of grammatical features), lemmata and superlemmata. The introduction of the superlemma level was particularly important for preserving the considerable orthographical richness of Medieval Latin as a historical language. This approach is analogous to the Wiktionary model of lexical units, but in contrast to Wiktionaries and above all wordnets (such as the Latin WordNet – cf. Franzini *et al.*, 2019) it lacks lexical-semantic relations (see Section 1).

The superlemma provides the normalized spelling of a lemma so that on the lemma level different spellings can be kept. The *FLL* currently contains 116,297 superlemmata and 133,691 subordinated lemmata. These lemmata have been expanded morphologically according to the standard grammar of Classical Latin as described in Menge (2009) and Rubenbauer *et al.* (2009) so that the *FLL* now has 9,663,808 syntactic words (see Table 1 for earlier statistics of the *FLL*)[16].

[7] For the following list see Mehler *et al.* (2015); see also vor der Brück and Mehler (2016) for more information about the *FLL*. The presentation of the *FLL* in this article is a correction of its earlier presentation in Mehler *et al.* (2015), which contained much higher amounts of overgeneration.

[8] Apparently, this resource no longer exists.

[9] Cf. *http://archives.nd.edu/words.html*; today *http://www.latin-dictionary.net/*.

[10] Cf. *http://www.corpusthomisticum.org/it/index.age*.

[11] Cf. *http://www.neulatein.de/*.

[12] Cf. *https://la.wiktionary.org/wiki/Victionarium:Pagina_prima*.

[13] This data was provided by Michael Trauth, Trier University.

[14] Cf. *http://www.neg.uni-tuebingen.de/*.

[15] This concerns mainly texts provided by the project Corpus Corporum of Philipp Roelli in Zurich (*http://www. mlat.uzh.ch/MLS/*), the Monumenta Germaniae Historica (*https://www.dmgh.de/*) and the Institut de recherches d'histoire des textes (*IRHT*; *https://www.irht.cnrs.fr/*).

[16] Apart from some exceptions like the oblique case, the grammar rules did not alter between Classical and Medieval Latin; cf. (Menge, 2009; Rubenbauer *et al.*, 2009).

PoS	Superlemma	Lemma	Syntactic Word	Description
ADJ	21,870	26,070	3,337,028	adjective
ADV	9,682	11,163	42,864	adverb
AP	86	117	482	preposition
CON	101	140	519	conjunction
DIST	46	49	1,321	distributive number
FM	76	109	2,343	foreign material
ITJ	112	115	254	interjection
NE	5,843	6,649	114,757	named entity
NN	35,433	45,383	745,345	common noun
NP	26,741	29,657	247,911	personal name
NUM	101	143	3,140	number
ORD	131	194	4,871	ordinal number
PRO	125	172	8,041	pronoun
PTC	12	17	38	particle
V	9,081	13,164	5,135,824	verb
XY	114	117	718	unknown
Sum	109,554	133,259	9,645,456	

Table 1. *Statistics of the* FLL *(release as of May, 2019): superlemmas, lemmas and syntactic words are listed together with their numbers and differentiated by 15 parts of speech, supplemented by a class of words (denoted by* XY*), which collects unknown cases.*

In the future, the Superlemma-ID will be used to connect the *FLL* with other lexical resources on the web (and also with resources provided by traditional long-term institutions for Latin lexicography), so that morphological information will be available together with reading aids. The lexicon could also work with a fourth ('lexeme group') and a fifth level ('synset') to bundle superlemmata of different PoS that share the same root, or to map semantic relations. But this is future work.

The *FLL* can map multi-word units, which makes it easier to record proper names such as *Colonia Agrippina*. However, the four-step model currently meets the requirements of lemmatization. The lexicon is managed via the Lexicon Browser of the eHumanities Desktop, which has been especially adapted for humanities scholars without programming skills. Entries can be created, changed, merged, reorganized or deleted by authorized users. The so-called Extension Tool then creates all inflected forms for newly entered lemmata.

Only basic information is needed to identify the right declination or conjugation for the new token. All changes are documented by naming the authors and timestamps. Additional columns allow descriptions to be added to the entries or to show if an entry has been double-checked. This procedure was developed together with various third-party funded humanities projects that have based their philological and linguistic research on their work with the *FLL*[17].

An objection to the method of morphological expansion, and thus to lexica of the type of the *FLL*, is to say that it is prone to overgeneration. To calculate this, we used as a reference corpus a repository of Latin texts including Migne Patrologia Latina (*MPL*), substantial parts of the Monumenta Germaniae Historica (*MGH*) and other repositories[18] to ask for the number of syntactic words of the *FLL* that our lemmatization finds manifested in this repository. This is shown in Table 2. Indeed, only 9% of the syntactic words of the *FLL* are found in this 'text repository'. Figure 1 shows how this coverage grows with the percentage of tokens of the reference corpus covered by the *FLL*.

Attribute	Value
All texts	111,515
All tokens	185,808,777
Tokens mapped to the *FLL*	180,535,369 (97.16%)
Tokens unassigned	5,273,408
All superlemmata in the *FLL*	109,554
Superlemmata used	83,780 (76.47%)
All lemmata in the *FLL*	133,259
Lemmata used	102,728 (77.09%)
All syntactic words	9,645,456
Syntactic words used	871,452 (9.04%)

Table 2. *On overgeneration and underrepresentation as induced by the* FLL *(release as of May, 2019).*

[17] See below.

[18] For the Monumenta Germaniae Historica (*MGH*) see the open*MGH* repository (*http://www.mgh.de/dmgh/openmgh/*); Migne Patrologia Latina (*MPL*) is available from the Corpus Corporum website (*http://www.mlat. uzh.ch/MLS/*); in addition the repository includes the Roman Law Library (*https://droitromain.univ-grenoble-alpes.fr/*), the corpus of Cluny Charters (*http://www.cb-ma-project.eu/*), parts from the Latin Library (*http://thelatinlibrary.com/*) and from the Central European Medieval Texts Series (*http://ceupress.com/series/ central-european-medieval-texts*).

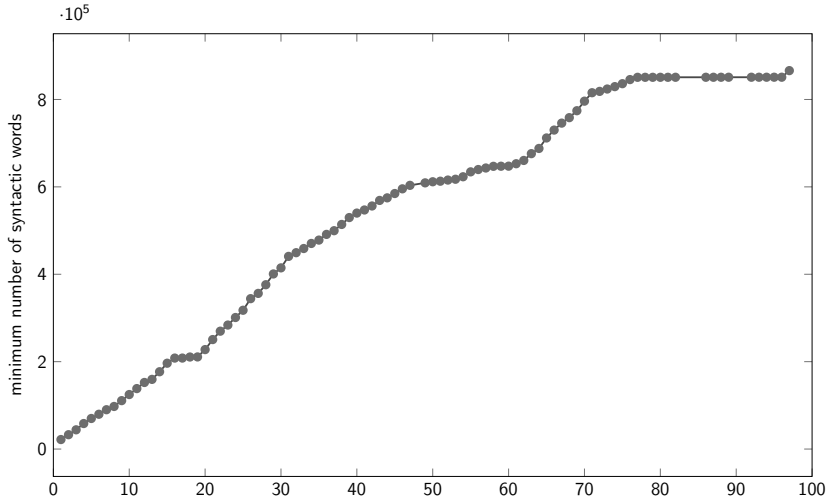

Figure 1. *The minimum number of syntactic words of the* FLL *(release as of May, 2019) (y-axis) sufficient to cover the corresponding percentage of tokens of the reference corpus.*

It demonstrates almost a linear trend – except for the rightmost part of the distribution: the number of syntactic words in the *FLL* that are observed in the corpus grows linearly with the size of this corpus. The coverages are much higher on the level of lemmata (77.09%) and superlemmata (76.47%). On the other hand, the underrepresentation, that is, the number of tokens within our reference corpus that are unknown from the point of view of the *FLL*, is remarkably low (2.84%). Such a rate of coverage did not seem to be possible in the early days of the *FLL*: in fact, it was the morphological expansion that made it an extensive lexicon that allows such rates, so that with each lemmatized token the associated grammatical features can be linked. Furthermore, the corpus of Medieval Latin texts within the eHumanities Desktop[19] is continuously extended, with each syntactic word being identified by a corresponding corpus frequency. These frequencies allow the subsequent filtering of supposedly overgenerated words for downstream tasks or even for information retrieval by users.

[19] It currently manages 112,657 Latin documents of different sizes (release as of November 2019).

3. *Crowdsourcing the* FLL

The *FLL* grows as new texts are uploaded into the text database of *HSCM*[20] and lemmatized. The result of automatic lemmatization is checked mainly by Latin philologists during the post-lemmatization process. They correct unfitting assignments between text and lexicon or create new entries in the *FLL* to close gaps in the lexicon and in the lemmatization. To this end, human editors can use the so-called Lemmatization Editor of the eHumanities Desktop. It presents lemmatized text in a color code and a statistical overview indicating the state of lemmatization. The color code differentiates nine distinct levels of lemmatization. The code does not only lead the human editor to tokens that still need to be identified or disambiguated but also marks lemmatization results with different degrees of certainty. The expert then opens the so-called Word-Link-Editor for a token to check the tagger's choice that the expert can confirm or correct. Here, she or he can disambiguate the result if the tagger has not taken a decision. In very few cases – concerning mainly proper names, OCR mistakes or abbreviated wordforms – the color code displays the token in blue which means that no corresponding entry could be found in the *FLL*. In this case, the human editor can either correct the misspelling directly within the Lemmatization Editor or create a new entry. If necessary, the expert can leave the editor and open the Lexicon Browser to create a new superlemma and/or expand lemmata. All these actions influence the state of lemmatization directly which becomes visible through the changing color code.

Evidently, the manual post-lemmatization process may detect errors in the *FLL*. In this case, the expert must correct the corresponding entries, merge duplicates, or delete incorrect entries. Such errors are mainly due to the initial setup when information was taken from different sources or through overgeneration as a result of morphological expansion. But even human editors sometimes make mistakes. Therefore the lexicon offers the possibility to mark entries as 'double-checked'. Since changes in the lexicon cannot simply be undone, changing the entries requires a high level of expertise, which is why the lemmatization and subsequent lexicon work is only carried out by trustworthy project members. As a result, this work is done using a limited crowdsourcing approach by assigning update rights to a limited

[20] *HSCM* stands for Historical Semantics Corpus Management, a system for managing the Latin text database of the eHumanities Desktop.

number of experts using eHumanities Desktop's rights management tool (Gleim *et al.*, 2012). In the project Computational Historical Semantics[21], linguists from the universities of Bielefeld, Regensburg and Tübingen, who worked directly with the *TTLab* and the Historical Seminar of the Goethe University, were allowed to update the lexicon. In addition, external partner projects from the Universities of Cologne, Freiburg and Mainz participate in the update process of the *FLL* after appropriate training[22]. As of May 2019, the percentage of lemmata created or modified by this procedure was 13.93% (18,565 lemmata).

The synchronization between the lemmatized corpus and the *FLL* as induced by post-lemmatization and lexicon updates is managed by *TEILex*, a system for integrating lexica and text corpora, in which the tokens of a corpus are linked with corresponding lexicon entries in such a way that lexicon updates are immediately transferred to the linked corpora and vice versa. In this way, expert-based lexicon modeling becomes less dependent of indexing the underlying corpus. Figure 2 shows the corresponding workflow of *TEILex*: automatic text processing using TextImager (which generates XMI files) and subsequent TEI conversion generates a TEI-conform corpus that is indexed and synchronized with the *FLL* using *TEILex*. In this process, experts can change both the lexicon (see update (1) in Figure 2) and the tagged corpus (update (3)). However, in order to prevent the corpus from being re-indexed after each change, the synchronization of the *TEILex* corpus with the *FLL* allows the automatic execution of changes of the lexicon on the synchronized corpora. The *TEILex* index is then automatically revised without having to interrupt the use of the system. This procedure has greatly accelerated the post-correction process by protecting it from too many interruptions.

[21] Cf. *www.comphistsem.org*.

[22] This concerns e.g. the project HUMANIST (2017-2020 at the universities Darmstadt, Mainz and at the Hochschule Mainz; *https://humanist.hs-mainz.de/*, funded by the Federal Ministry of Education and Research). In this context, a current project at Johann Gutenberg University is establishing a digital version of the so-called various letters written by the eminent 6th century politician and philosopher Cassiodorus (d. ca. 585). Here, specialists are checking the automatic lemmatization provided by means of the *FLL* and produce completely disambiguated texts for their project's purposes. Particularly worth emphasizing is how they make their work transparent – see *https://humanist.hs-mainz.de/projekt/inhaltlicher-projektkern/digitale-edition/*; a Freiburg University based partner project, funded by the German Research Foundation, focused on high medieval feudal law and imperial charters (*https://gepris.dfg.de/gepris/projekt/264932155*); another partner project focusses on 6th to 9th century Frankish royal edicts, so called capitularies (*https://capitularia.uni-koeln.de/*, a long term project funded by the Union of the German Academies of Sciences and Humanities).

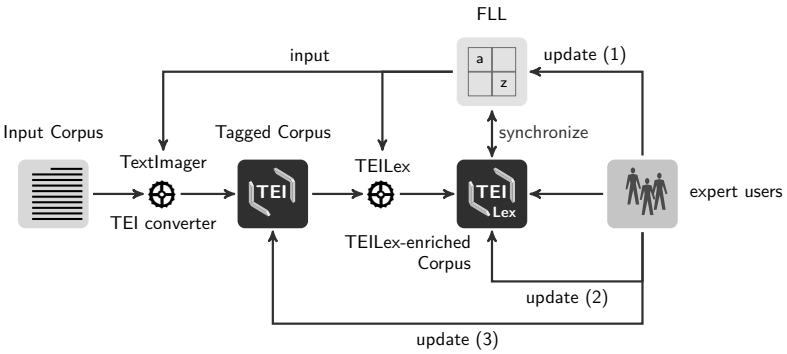

Figure 2. *Schematic depiction of* TEILex' *workflow.*

4. *Lemmatization in the flux*

In Eger *et al.* (2016) and Gleim *et al.* (2019), we experimented with different lemmatizers on Latin datasets. Since the publication, thanks to the emergence of large transformer networks (e.g. *BERT*; Devlin *et al.*, 2019), significantly better models have appeared, which prompted us to update our results. These transformer networks are large language models that are trained on even larger amounts of data and recognize and process syntactic and semantic relations (Clark *et al.*, 2019). These attention-based networks with up to 24 layers and over 350 million parameters (sometimes even more; see Shoeybi *et al.*, 2019) are trained on natural language texts (mostly Wikipedia and digital books) with the task of reconstructing deleted words by their context. These models can then be adapted to specific tasks on the basis of training data (Kondratyuk and Straka, 2019). *BERT* (Devlin *et al.*, 2019) is still the most popular model, as it was pre-trained in 104 languages and is publicly available. In addition, the models are not too large, so that they can be (post-)trained for individual purposes without the need for special hardware. The most advanced lemmatization model currently available from this class of approaches is *UDify* (Kondratyuk and Straka, 2019), which is based on a multilingual *BERT* model. *UDify* has been fine-tuned on 124 treebanks in 75 languages and is capable of tagging universal PoS[23], morphological features, lemmata, and dependency trees and also obtains acceptable results in

[23] For a recently published study on PoS tagging in Latin see Stoeckel *et al.* (2020).

unknown languages. Under these 124 treebanks are 3 in Latin: *PROIEL* (Haug and Jøhndal, 2008), Perseus (Bamman and Crane, 2011) and *ITTB* (Cecchini *et al.*, 2018; Passarotti and Dell'Orletta, 2010) with a total of 582,336 tokens. The *PROIEL* treebank contains most of the Vulgate New Testament translations plus selections from Caesar and Cicero. *ITTB* (i.e. the Index Thomisticus Treebank) contains the complete work by Thomas Aquinas (1225-1274; Medieval Latin) and by 61 other authors related to Thomas, while Perseus contains a selection of passages from diverse authors like Augustus and Tacitus. All treebanks are annotated with the Universal Dependencies[24] (*UD*) Framework (Nivre *et al.*, 2016). In order to adapt *UDify*, the texts are preprocessed with the help of *BERT*, whereby for each downstream task a separate classifier is trained with the help of the resulting *BERT* word vectors. The *UDPipe* model (Straka and Straková, 2017), on the other hand, is a pipeline system (i.e. a system that interconnects series of NLP tools) that does not rely on large transformer models, but is designed independently for each target language and therefore cannot (directly) exploit similarities between languages. As a consequence, 97 independent models were trained on 97 treebanks from 64 languages (Straka and Straková, 2017).

We have tested both models on our corpus of Frankish royal edicts, the capitularies (Mehler *et al.*, 2015), and on the *PROIEL* corpus (Haug and Jøhndal, 2008) and compared the results with the taggers from our original work (Gleim *et al.*, 2019). All results are listed in Table 3.

First we concentrate on the results on the *PROIEL* corpus. *UDPipe* achieves significantly better results on this dataset than *UDify*, although both were trained on this dataset. Generally with 96.32% a very good F1 score[25] is achieved by *UDPipe*. However, the dataset used by *UDify* for training also included the *ITTB* and Perseus data. It is not surprising that the tools that were trained on the Capitularies perform significantly worse on the *PROIEL* data. The evaluation on the Capitularies, on the other hand, is more interesting since neither *UDPipe* nor *UDify* were trained on it. *LemmaTag*, which was also trained on the Capitularies (Gleim *et al.*, 2019), reaches an F1 score of more than 96%. Taggers such as *LemmaGen*, *MarMoT* and *LemmaTag* on the other hand, which were only trained on *PROIEL*, generalize much worse when being evaluated out-domain by means of the Capitularies; this can be an indicator of overfitting. *UDify*, which was trained

[24] Cf. *https://universaldependencies.org/*.

[25] The F1 score is the harmonic mean of precision and recall of the corresponding classification.

Lemmatizer	Trainings Corpus	*PROIEL* (Haug and Jøhndal, 2008)	Capitularies (Mehler *et al.*, 2015)
LemmaGen (Juršic *et al.*, 2010)	Capitularies (Mehler *et al.*, 2015)	81.39 (Gleim *et al.*, 2019)	95.64 (Gleim *et al.*, 2019)
MarMoT (Müller *et al.*, 2013)	Capitularies (Mehler *et al.*, 2015)	81.24 (Gleim *et al.*, 2019)	95.81 (Gleim *et al.*, 2019)
LemmaTag (Kondratyuk *et al.*, 2018)	Capitularies (Mehler *et al.*, 2015)	82.25 (Gleim *et al.*, 2019)	**96.13** (Gleim *et al.*, 2019)
LemmaGen (Juršic *et al.*, 2010)	*PROIEL* (Haug and Jøhndal, 2008)	90.63 (Gleim *et al.*, 2019)	76.28 (Gleim *et al.*, 2019)
MarMoT (Müller *et al.*, 2013)	*PROIEL* (Haug and Jøhndal, 2008)	90.29 (Gleim *et al.*, 2019)	76.37 (Gleim *et al.*, 2019)
LemmaTag (Kondratyuk *et al.*, 2018)	*PROIEL* (Haug and Jøhndal, 2008)	81.85 (Gleim *et al.*, 2019)	49.61 (Gleim *et al.*, 2019)
UDPipe (Straka and Straková, 2017)	*ITTB* (Passarotti and Dell'Orletta, 2010)	---	83.80
UDPipe (Straka and Straková, 2017)	Perseus (Bamman and Crane, 2011)	---	78.87
UDPipe (Straka and Straková, 2017)	*PROIEL* (Haug and Jøhndal, 2008)	**96.32** (Kondratyuk and Straka, 2019)	86.94
UDify (Kondratyuk and Straka, 2019)	124 treebanks	91.79 (Kondratyuk and Straka, 2019)	88.25

Table 3. *Results of lemmatizers trained on different data (rows) and evaluated on the* PROIEL *corpus and our Capitularies (Frankish royal edicts, 6th to 9th c.) corpus (columns). Bold indicates best results of in-domain and underlined of out-domain lemmatization. The references behind the results refer to the paper in which they were published. Unreferenced scores indicate newly trained models.*

on three Latin corpora and several other languages, generalizes much better: being evaluated out-domain by means of the Capitularies, it still reaches an F1 score of 88.25%. Among all taggers which were not trained on the Capitularies, *UDify* achieves the best results. This ability to generalize makes it a very interesting candidate for lemmatization. The results between the corpora may have been even stronger, but there are differences in annotation between them. This is particularly evident in the errors that *UDify* makes most often (as listed in Table 4). Just by fixing these errors by means of a simple post-processor, *UDify*'s performance can be considerably improved.

Form	Gold	Predicted	Count
a	*a*	*ab*	2020
se	*sui*	*se*	999
quod	*quod*	*qui*	892
ac	*ac*	*atque*	884
sibi	*sui*	*se*	371
vero	*vero*	*verus*	345
seu	*seu*	*sive*	342

Table 4. *Most frequent errors made by* UDify *on the Capitularies.*

This analysis shows that in-domain lemmatization can be delegated to modern neural network models that appear to be largely independent of lexicons of the type of the *FLL*. Even those models of Gleim *et al.* (2019), which use the *FLL*, are outperformed by transformer-based models in the area of out-domain lemmatization. From the point of view of the manual post-lemmatization process, however, the reference to a lexicon remains indispensable when it comes to distinguishing between lemmata and superlemmata and correctly assigning them to incorrectly lemmatized tokens – a residual task that can probably never be fully automated. That is, regardless of the enormous progress achieved by transformer-based taggers, an out-domain F-score of 88% (as demonstrated by *UDify* on the Capitularies) falls short of the threshold that would be acceptable from the point of view of humanities scholars. And even if one assumes that F-scores around 98% are practically unattainable, since inter-rater agreements also fall short of this margin, the requirement remains for human post-processing and especially post-lemmatization, which requires corresponding lexicon-based guidance as addressed, for example, by the *FLL*. And since a lexicon such as the *FLL* requires inte-

gration with distributional semantic resources sprouting up everywhere, an answer is needed to the question of how the compact vector representations of such approaches can be mapped to manageable graphs, which in turn can be consulted by expert users to refine their lexicon work. An answer to this question will be sketched in the following section.

5. *Genre-sensitive embeddings in Latin*

Gleim *et al.* (2019) show that using word embeddings can boost lemmatization also in Latin to a remarkable degree. However, the embeddings involved are calculated for large corpora, such as the Patrologia Latina (Jordan, 1995), without taking, for example, the genre-related diversity of text vocabularies into account: as with lexical ambiguities, such embeddings model varieties using composite structures without providing separate representations for them. Apparently, approaches of this sort assume that resources are homogeneous data: they operate on as much data as possible from genres, registers or time periods that do not exhibit substantial heterogeneity or whose heterogeneity is ignored by the model. In this article, we take a different approach: by subdividing corpora according to their contextual stratification, we obtain subcorpora for training specialized embeddings that differentiate knowledge which is otherwise amalgamated within the same model. This makes it possible to explicitly represent differences of the same word due to its varying use in different genres, subject areas or stylistic contexts. In this way, reference is made to linguistic knowledge in order to make the computation of lexical relations more transparent. The further goal is to improve the interpretability of machine learned resources from the point of view of the targeted community.

Thus, in light of Section 1, our goal is to extend *FLL* so that for each word a series of embeddings is learned that are differentiated according to a subset of contextual dimensions (e.g. author, genre, style, register, topic, etc.). *FLL* then no longer represents words (superlemmata, lemmata, wordforms or syntactic words) as nodes of a monoplex network, but as nodes of a multiplex network (Boccaletti *et al.*, 2014) whose multiplexity is established by context dimensions. Henceforth, we denote this variant of *FLL* by *FLL+*: *FLL+* is a terminological ontology that spans a multiplex network according to different contextual dimensions and thus provides a series of contextualized representations for each of its lexical entries – as (downloadable) embeddings and as traversable SemioGraphs (see below). Multiplicity means to

network the vertices of the same graph according to different (in the present article: contextual, discourse-level) criteria. Furthermore, the embeddings used to network *FLL+* as a multiplex network are partly hierarchically ordered by the subset relations of the corpora involved. This means, among other things, that word embeddings calculated for a subcorpus *x* of a corpus *y* can be used to approximate embeddings calculated on the latter.

In order to generate *FLL+*, we consider genre as a contextual dimension by example of six instances (see Table 5): 'epistolographic', 'legal,' 'liturgical', 'narrative', 'political' and 'theological' texts.

	#Text	#Tokens
Reference corpus	111,515	185,808,777
Overall training corpus	33,791	61,451,677
Epistolographic texts	844	16,406,556
Legal texts	31,461	12,097,990
Liturgical texts	252	2,667,784
Narrative texts	663	7,635,906
Political texts	31	3,197,879
Theological texts	494	18,305,475

Table 5. *Statistics of the corpora used for computing specialized embeddings.*

Further, we analyze authorship as a contextual dimension by example of three authors: Bernard of Clairvaux (d. 1153), John of Salisbury (d. 1180) and William of Ockham (d. 1347). Last but not least, we compute embeddings for our reference corpus of 111,515 texts[26]. This corpus contains the Patrologia Latina, historiographical and legal texts from the Monumenta Germaniae Historica and additional historiographical texts from the Central European Medieval Texts series. The corpus can be accessed by means of the eHumanities Desktop. The same applies to the special corpus of legal texts analyzed here (see Table 5) that contains the *Corpus Iuris Civilis* (compiled 528-534) and the *Corpus Iuris Canonici* (gradually compiled from the mid-12th to the 15th century) next to canonical decrees and Carolingian law texts. This approach of contrasting the reference corpus with specialized subcorpora makes it pos-

[26] All embeddings are available for download at *http://embeddings.texttechnologylab.org*.

sible to compare embeddings generated by means of the reference with those obtained for specialized genres (see Section 6 for such a comparison).

Since our focus is on genre and author-related variation and not on method optimization, we concentrate on efficiently computable methods:

(i) We utilize the well-known continuous bag-of-words (*CBOW*) and the skip-gram model of word2vec (Mikolov *et al.*, 2013).

(ii) As a further development of word2vec, we experiment with fastText (Joulin *et al.*, 2017), which additionally evaluates character embeddings for computing word embeddings. This approach again comes in two variants: skip-gram and *CBOW*.

(iii) As we deal with subsets of corpora of varying size (see Table 5), we also experiment with a method that addresses 'small' corpora. This relates to the approach of Jiang *et al.* (2018), who evaluate the common absence of words in text segments as an additional source of information (cf. Rieger, 1989). An alternative approach is the one of Silva and Amarathunga (2019), who generate random paths in sentence networks to obtain sentence variants for extending small input corpora. Both approaches have been evaluated in the context of word similarity tasks and are promising candidates for dealing with low-resource situations. However, we concentrate on the approach of Jiang *et al.* (2018).

Starting from the resulting embeddings and their specialization for different genres, we generate so called 'local graph views': instead of comparing embedding representations as a whole (cf. Veremyev *et al.*, 2019; Yaskevich *et al.*, 2019)[27], local views generate local neighborhoods of words. For this purpose we generate for all words of the *FLL* the graph induced by their 100 nearest neighbors. In this way, we get for each word of the *FLL* 50 = 10 (1 reference corpus + 9 subcorpora) × 5 (computational methods) different embeddings. Since the *FLL* distinguishes between wordforms, syntactic words, lemmata and superlemmata, this finally multiplies to 200 different embeddings to be managed[28].

Since our goal is not only to generate distributional semantic resources for Latin text genres, but to make them also interactively available to the (expert) user, we finally generate so-called SemioGraphs[29] as visualizations of local graph views using the pipeline[30] of Figure 3.

[27] Cf. also *https://github.com/anvaka/word2vec-graph*.

[28] For reasons of space complexity we compute only a subset thereof.

[29] Cf. *http://semiograph.texttechnologylab.org/*.

[30] This pipeline is available at *https://github.com/texttechnologylab/SemioGraph*.

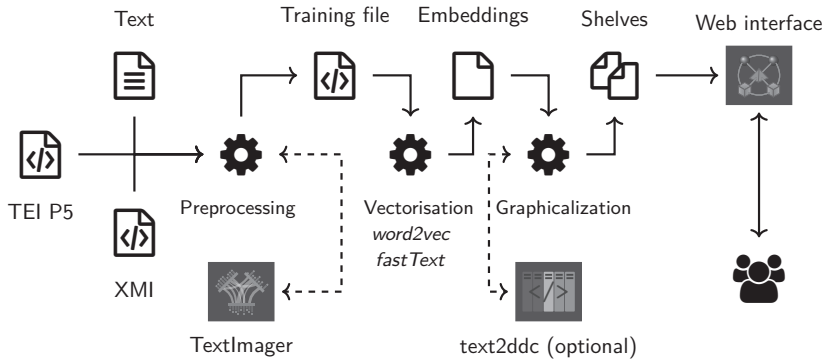

Figure 3. *Processing pipeline for generating local graph views from embeddings possibly enhanced by topic labels as input to interactive traversable SemioGraphs.*

It includes three steps to process plain text, TEI or XMI[31] documents where TextImager (Hemati *et al.*, 2016) is used to preprocess input documents based on the procedures for processing Latin documents described in Gleim *et al.* (2019): preprocessing creates training files from single documents or entire repositories as input for vectorization (e.g. by means of word2vec etc.). The third step ('graphicalization') generates input files required by the SemioGraph application, a Python script that uses the flask server framework, which is available as a Docker image for rapid setup. The SemioGraph application manages all generated embeddings for generating graph views. The output of graphicalization is stored in Python shelve files, that is, key value stores of Python objects that allow for fast access by the server. If available, graphicalization enriches SemioGraphs on the node level with topic labels using text2ddc (Uslu *et al.*, 2019), which is currently trained for 40 languages (but not yet for Latin). The Python shelves are finally used to generate interactive visualizations using SemioGraph's web interface. In this interface, node size codes two dimensions of vertex-related salience: while 'height' codes degree centrality, 'width' is used to code the similarity to the seed word. Furthermore, 'node transparency' can be used to code degrees of class membership values, while 'node color' can map the corresponding class (this feature is not used in our example below). Beyond that, 'border color' can be used to code a 2nd-level vertex-related classification (e.g. topic-related class membership). Finally,

[31] XMI is a serialization format for UIMA-CAS documents.

in the case of multilabel classifications, 'node tiling' (i.e. pie charts) can be used to code distributions of class membership per vertex (also this feature is currently not used by our Latin SemioGraphs).

Generally speaking, a SemioGraph is a visual, interactive representation of word embeddings as a result of the latter procedure: starting from a word x, its SemioGraph displays those m words which by their embeddings are among the first m neighbors of x in the similarity space induced by the underlying embeddings. This means that vertices or nodes in a SemioGraph represent words or multi-word units, while edges or links represent associations of these nodes, the strength of which is represented by the thickness of the (visual representation of the) edges. Since word embeddings induce fully connected graphs (in which all words are connected with each other), the SemioGraph interface allows to filter low associations to get visual insights into the underlying graph structures: this enables the visual formation of clusters of nodes, which have a higher number of internal associations than to members of other clusters or to outliers. Thus, if a SemioGraph of a word x is generated using this method, this does not mean that visually disconnected words are not associated. Actually, they are, but to a degree below the user-controlled threshold value. In any case, due to the way embeddings are calculated here (if being based on the *CBOW* model as done below), SemioGraphs show paradigmatic associations. This means that even if word co-occurrences are frequent (indicating higher syntagmatic associations), the word associations need not appear in the corresponding SemioGraph. This happens in cases where the contexts in which the words are used throughout the underlying corpus are less similar than their inclusion among the m most similar neighbors would require: a SemioGraph always shows only a selected subset of associations. Thus, not appearing in a given SemioGraph does not necessarily mean non-existence. If the latter selection would include all words from the input corpus, then these would all be displayed in the SemioGraph, of course by means of edge representations of variable thickness. Visualizing genre-sensitive embeddings using SemioGraphs then means first generating word embeddings separately for corpora that reflect certain genre-, register-, chronology-, or other context-related features, and then visualizing the neighborhoods of certain seed words to determine the differences or similarities of their context-sensitive syntagmatic or paradigmatic associations. This is illustrated in detail in the next section by means of paradigmatic word associations.

6. *Brief case studies in computational historical semantics with the help of SemioGraph*

In this section, we apply the method of local graph views to Latin word embeddings as provided by SemioGraph, briefly present four SemioGraphs and sketch how they may provide a new kind of evidence for computational historical semantics in the humanities. In our first example we calculate paradigmatic associations (Rieger, 1989) of the noun *conclusio* "conclusion" in the test corpus of legal texts (see Table 5). The resulting SemioGraph (see Figure 4) allows first observations on the principle functionality of SemioGraphs for a comparatively small corpus, on the potentials of genre-sensitive SemioGraphs, and at the same time on necessary further work and current performance of the *FLL* and *TTLab*'s tagger and lemmatizer for Latin.

Figure 4. *Local graph view of* conclusio *(*NN*); genre: legal texts (see Table 5); method:* CBOW *(Mikolov* et al.*, 2013).*

On the one hand, the calculated semantic connections in Figure 4 correspond to what for historians fits very well into a well-known context of legal history. First, there are mostly technical terms for different aspects in legal

processes – "inquiry", "examination", "excuse", "allegation" (*cognitio*, *examinatio*, *excusatio*, *denuntiatio*, etc.) – from the dispute to its settlement. Time expressions such as "ten-year" (*decennius*) or "four-year" (*quadriennium*) may not necessarily refer to the length of punishments, but rather to time specifications in legislation. Secondly, there is a single recognizable content, that is, marriage legislation, and this is quite clearly visible. The graph shows a vocabulary that historians would expect in texts on marriage legislation of these centuries – "divorce" / "repudiation" (*divortium*, *repudium*), "copulation" or "copulate" (the term *copulam* signals the need for manual post-lemmatization), the legal importance of the 'consumption' of a marriage (*consummatio*), "puberty" (*pubertas*), the "conjugal union" (*matrimonialis* – adjective), or "conceiving" and "childbearing" (*partum*, specification via post-lemmatization is needed).

On the other hand, links between many words – as well as the occurrence of words without any links – indicate a need for further clarification, which must be systematized in the workflow of such queries: no link or edge connects the keyword *conclusio* with any of the other words. Apparently, *conclusio* does not co-occur in the underlying corpus with any of these words with sufficient frequency. Links between words require a certain minimum number of neighborhoods, which serve as a reliable source of information for their paradigmatic associations. Note again that if a SemioGraph shows no link between two words, this does not mean that they are not related to each other; it only means that their paradigmatic association is below a certain minimum, where the user of the SemioGraph sets this threshold him- or herself.

Some of the key words of marriage legislation such as 'copulation' or 'consumption' are also disconnected in the SemioGraph in Figure 4. Some of these observations may disappear with the enlargement of the underlying corpus by means of texts that provide more evidence about their contextual similarities (Miller and Charles, 1991). In any case, the calculation of paradigmatic associations ultimately aims at making such phenomena visible. That is, associations should become visible even if the words involved are rare in the underlying corpus, but the similarities of the contexts in which they are used are sufficiently strongly confirmed by that corpus. It is therefore less a matter of eliminating such observations in a SemioGraph (in terms of post-correction) than of making them (i) controllable by means of corpus selection and (ii) interpretable with respect to this selection. One of several possible explanations may be that the keyword has not been used in standardized collocations: and such an observation can then be the starting point for research in the respective humanities.

In the second example (see Figure 5) we look for the 50 closest neighbors of the verb *excommunico* ("I exclude someone from the ecclesiastical community") in our corpus of legal texts.

Figure 5. *Local graph view of* excommunico *(*V*); genre: legal texts; method:* CBOW *(Mikolov* et al.*, 2013).*

Since excommunication was one of the few punishments the church could impose on someone, the verb can be expected mainly in legal contexts. Again, we find a semantic field that includes many tokens a historian would expect. To discuss some of them: first, *excommunico* (as well as *excommunicatio* – noun, and for post-lemmatization the spelling variations: *excomunico*, *excomunicatio*) appears in the SemioGraph together with the terms for "anatheme" (*anathema* – noun, *anathematicus* – adjective, *anathemizo* – verb, *anathematizatus* – participle) and for "interdict" (*interdictio* – noun, *interdico* – verb, *interdictus* – participle). These three legal terms are – following common encyclopedias and dictionaries such as the *Lexikon des Mittelalters* (Zapp, 1980: 574; 1989: 170; 1991: 466-467) – difficult to distinguish, they were used interchangeably for describing very similar situations. Facing the current status of our genre-specific corpus of legal texts, which is still prelim-

inary, the SemioGraph gives the impression that forms of *anathema* and of *excommunico* do indeed have a fairly similar set of neighboring words (in the sense that the SemioGraph displays many shared links). At the same time, we observe missing edges or links visualized by thin lines. This in turn indicates that the words concerned are associated below the threshold for visualizing such relations. Considering phrases like "excommunicate or anathemize" (*excommunicare vel anathematizare*) or "suspend or excommunicate" (*suspendere vel excommunicare*) used by some very influential authors (Regino of Prüm, Burchard of Worms, Ivo of Chartres), one may wonder why these co-occurrences are not reflected in the graph among the 50 closest neighbors[32].

The same observation can be made with *interdictum*, the third central weapon for hard ecclesiastical punishment. This term shows even fewer links to *excommunico* in the SemioGraph than the forms of anatheme, although again more than 200 times *excommunico* and a form of *interdictum* co-occur in sentences of the underlying corpus. A check in the corpus shows that, although pairings such as those cited may appear sufficiently frequent overall, the individual pairings are actually not sufficiently frequent to cross the threshold. These kinds of observations lead to further questions, especially what kind of calculation – by sentences or by word distances – brings the graph closer to the notion of 'paradigmatic associations', and how the observation of paradigmatic neighborhoods relates to classical co-occurrence analyses. Other terms in the SemioGraph of Figure 5 express reasons for excommunication like "heresy" (*haeresis* – noun), "disobedience" (*inobedio* – verb), "contumacy" (*contumax* – adjective) or "rebellion" (*rebellis* – adjective), as well as for the lifting of the excommunication (central: *reconcilio*; among the probable terms that are according to the SemioGraph not used as neighbors of *excommunico*: *absolvo*, i.e. "absolve"). Astonishingly, we do not find expressions for the holy community of the church itself among the paradigmatic neighbors. This also would demand further investigation.

In a third example, we look for the 50 closest associations of "father", in Latin *pater*, first within the complete reference corpus (a mere repository from patristics to the 15th century, Figure 6) and then in our legal texts corpus (Figure 7).

[32] One may also ask whether the SemioGraph can be associated more with methodological questions of computing rather than with historical phenomena. The reason is that any calculation of word association requires the fixation of certain parameters such as the number of neighbors in a sentence or the length of sentences in which neighbors are observed (and this holds of course also for SemioGraph). Any such parameter setting carries the risk of excluding relevant contexts – this is a general characteristic of computational linguistic analyses.

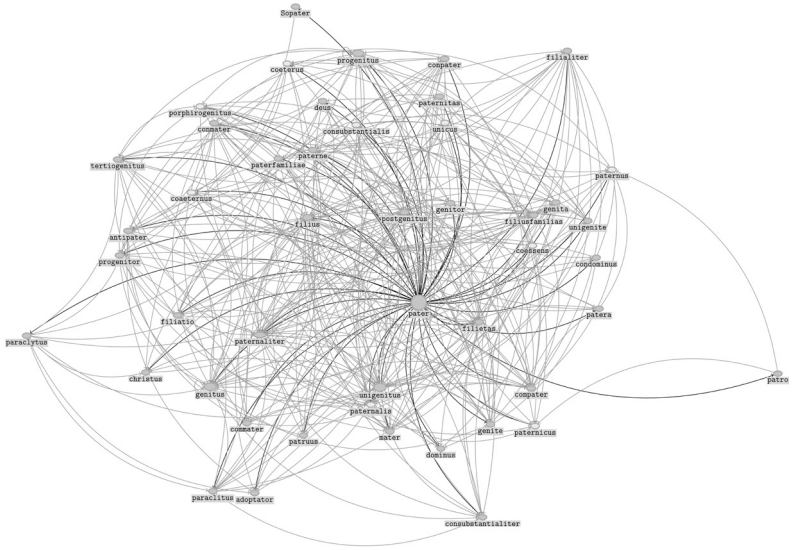

Figure 6. *Local graph view of* pater *(*NN*) taken from the reference corpus; method:* CBOW *(Mikolov* et al.*, 2013).*

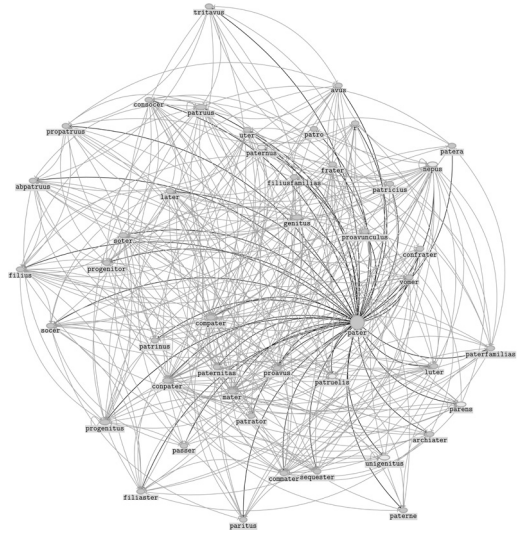

Figure 7. *Local graph view of* pater *(*NN*); genre: legal texts; method:* CBOW *(Mikolov* et al.*, 2013).*

This experimental arrangement follows the expectation that in the corpus of all texts we might see central religious aspects (Christ as the Son of God, Mary as the Mother of God), while in the corpus of legal texts we might see the Roman Catholic normative settings of the kinship system. The SemioGraph in Figure 6, which visualizes the results for our whole repository including all sorts of texts, shows some inconspicuous, culture-unspecific words such as "father" (*pater*), "mother" (*mater*), "son" (*filius*), "uncle" (*patruus*) or "paternal" (*paternalis*). We also find words that seem to be specific to the ancient Roman Mediterranean kinship system, such as *paterfamilias*, i.e. the father as head of a household, *materfamilias* or rarely *paterfamiliae* (the word *paterfamiliae* is sometimes used; however, the wordforms that have been automatically subsumed under this lemma mostly should be subsumed under *paterfamilias*). But a first test of the diachronic distribution suggests that they are not specific to a time. This observation reminds us that the repository brings together texts from two very different social systems – from the Greco-Roman Mediterranean ('Antiquity') and from the post-Roman Latin societies ('Middle Ages'). The semantics referring to the core Christian faith are clearly recognizable. The centrality of *unigenitus* in the SemioGraph points to "God's only begotten Son" (*unigenitus dei filius*, see also *deus* and *Christus* in the SemioGraph), an often repeated phrase of the Catholic creed. The strong connection of the word *unigenitus* to consubstantiality ("of the same substance", *consubstantialis*, *consubstantialiter*) indicated by the edges in the SemioGraph stresses the link to the Catholic creed. The SemioGraph reflects the prevailing religious attitude towards paternity bonds which subordinated carnal to spiritual paternity. Fatherhood of God, priests and godparents was a strong discursive element. The important role of godparents is visible by means of the terms "co-father" and "co-mother" (*commater*/*conmater* and *compater*/*conpater*) as neighbors of *pater*. Less visible are clerics as spiritual fathers since they were simply addressed as "father". Significant is also the lack of a broad family vocabulary that would differentiate family relationships. Only the mother and the paternal uncle (*patruus*) are present in this SemioGraph. This observation coincides at first glance with the broadly discussed hypothesis that the Roman male agnatic kinship system faded away under the influence of the church from the sixth century on in favor of kin groups organized around the conjugal couple (see Jussen, 2009). There is hardly any evidence of genealogical connections and far-reaching family relationships in this graph. This will be different in the SemioGraph in Figure 7.

Odd occurrences – such as the rarely used proper name *Sopater* as part of the SemioGraph in Figure 6 – immediately raise doubts and are hence particularly important terms for cross-checks of automatic procedures, that is, for intellectual post-correction. *Sopater*, correctly lemmatized in the *FLL* as a proper name, is an obvious candidate to check how paradigmatic similarities and corresponding neighborhoods are calculated. Even more conspicuous is the rather central position of the term *paterfamiliae*, a very rare variant of the common but here completely missing lemma *paterfamilias*. Both are subsumed under the same superlemma within *FLL*. In any event, two-dimensional geometric representations of graphs should not be overinterpreted – they may be due more to the visualization method and less to the underlying graph topology.

Such obvious but rare problems, however, are contrasted in Figure 7 by a multitude of plausible and in terms of interpretation controllable links, so that the added value of paradigmatic graphs for Latin texts can be regarded as successfully tested alongside the classical analyses of co-occurrences and syntagmatic patterns. First of all, it is striking that the kinship designations and the distinctions between the maternal and paternal lines, which were missing in the first graph, are prominent here (*avus* – "grandfather", *proavus* – "great-grandfather", *propatruus* – "great-granduncle", *abpatruus* – "great-great-great uncle from the father's side", a very rare word by the way, *tritavus* – "a grandfather's great-grandfather", *proavunculus* – "great uncle from the mother's side"). It is also striking that the designations almost exclusively refer to the paternal line. Since canon law has developed the kinship designations in both lineages in detail (in connection with the prohibition of incest), this finding again requires verification, that is, further research by the humanities scholar. In this case, the compilation of the corpus probably needs to be corrected. Presumably, charters function in linguistic terms differently from normative legal texts so that the corpus of legal texts should be divided into two corpora in future work. Furthermore, the data of the SemioGraph will probably only become meaningful when the long-term diachronic corpus of legal texts can be examined according to time sections. Only then will it be possible to see what was different in the Roman Mediterranean world compared to the post-Roman Latin-Christian societies.

These short case studies may suffice as an example for the implementation of computational tools like SemioGraph and the *FLL* in academic cultures with a very long hermeneutical tradition:

(i) The implementation of such tools in the humanities will have to start by overcoming mistrust. The results of the SemioGraphs must therefore be able to mirror the expectations and the 'assured knowledge' of the humanities in order to promote confidence in the reliability and controllability of the computerized calculation results.

(ii) Only then can they successfully manifest the unexpected that deviates from the discipline's 'assured knowledge'.

(iii) In this way, the SemioGraphs may motivate 're-reading', no longer guided by the authority of a very long hermeneutical tradition (which inevitably privileges certain canonized 'famous' texts), but stimulated by the authority of well controllable and comparable corpora.

(iv) Central to the acceptance of computational tools in the historical humanities is the strict and disciplined distinction between repositories and corpora, as we have shown in our last example.

It is these steps that must be achieved in order to institutionalize a lasting improvement of knowledge resources such as the *FLL* and SemioGraphs.

Establishing SemioGraphs as a tool for the visualization of paradigmatic associations in disciplines such as history or literary studies, theology or philosophy is no easy task. Despite all the changes that digitalization has brought with it, these disciplines will remain 'children of hermeneutics'. The success of any new methods depends on the ability to control the evidence in relatively small steps. The examples presented here can point to a way in this direction. In this article the focus was more on the technical possibilities, with some test cases as illustrations. It is left to a follow-up study to systematically verify the empirical gain – for example by examining one and the same seed word in all research perspectives mentioned here, that is, syntagmatic versus paradigmatic analyses, different definitions of neighborhoods (within one sentence, in the syntagmatic neighborhood of n words etc.), comparisons of different text types and different time layers of Latin texts (Roman world up to ca. Justinian, post-Roman Latin societies 6th-11th century, 12th-16th century). Only such a multi-perspective analysis can help to assess the added value and reliability of analyses such as those exemplified here.

By means of these case studies, we obtain an example of the triadic role of computational tools such as SemioGraph from the perspective of the applying humanities. That is, such tools serve:

(i) to meet and confirm the expectations of the scholars involved,

(ii) to manifest the unexpected that deviates from the current state of knowledge of the discipline, and

(iii) to motivate subsequent processes of 're-reading' in order to substantiate possible interpretations of the unexpected finding.

As far as this 'new reading' is equipped with tool chains of the kind outlined in this article, it could eventually lead to updates of the underlying knowledge resources, that is, the *FLL* and the embeddings based thereon, which in turn require updates of corresponding SemioGraphs, so that we finally get a manifestation of a digitally enhanced hermeneutic circle. We are convinced that it is worth pursuing this research direction further.

7. *Conclusion*

In this article, we presented the Frankfurt Latin Lexicon (*FLL*) as a dictionary resource for Latin that distinguishes between word forms, syntactic words, lemmata and superlemmata and thus implements a word model known from the Wiktionary project. We outlined a restricted crowdsourcing process by means of which the *FLL* is continuously checked and updated as well as the lemmatization of texts based thereon. We additionally reported progress in the lemmatization of Latin texts and stressed the need to enhance the *FLL* by means of word embeddings that are stratified according to contextual parameters such as genre, authorship and chronological order. Then, we introduced SemioGraphs as a means to interact with and traverse this embedding information. Finally, we presented case studies based on SemioGraphs using word embeddings computed for selected seed words of the *FLL*. Since these case studies show the need for downstream processes of close reading and possibly for corrections of the underlying lemmatization, we have identified in this process chain an instance for a 'digitally enhanced hermeneutic circle'. It could be seen as an example of a prototypical strategy for dealing with lemmatization or, more generally, natural language processing of historical language texts. Future work will focus on a more detailed examination of word embeddings in Latin, their local and global graph representations, and in particular on their intrinsic evaluation.

References

Bamman, D. and Crane, G. (2011), *The ancient Greek and Latin dependency treebanks*, in Sporleder, C., van den Bosch, A. and Zervanou, K. (2011, eds.), *Language Technology for Cultural Heritage*, Springer, Berlin, pp. 79-98.

Boccaletti, S., Bianconi, G., Criado, R., Del Genio, C.I., Gómez-Gardenes, J., Romance, M. Sendina-Nadal, I., Wang, Z. and Zanin, M. (2014), *The structure and dynamics of multilayer networks*, in «Physics Reports», 544, 1, pp. 1-122.

Cecchini, F.M., Passarotti, M., Marongiu, P. and Zeman, D. (2018), *Challenges in converting the Index Thomisticus Treebank into Universal Dependencies*, in de Marneffe, M-C., Lynn, T., Schuster, S. (2018, eds.), *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, Association for Computational Linguistics, Stroudsburg, PA, pp. 27-36.

Clark, K., Khandelwal, U., Levy, O. and Manning, C.D. (2019), *What does* BERT *look at? An analysis of* BERT*'s attention* [preprint available online at *https://arxiv.org/abs/1906*].

Crane, G. (1996), *Building a digital library: The Perseus project as a case study in the humanities*, in Fox, E.A. and Marchionini, G. (1996, eds.), *Proceedings of the First ACM International Conference on Digital Libraries, DL '96*, Association for Computing Machinery, New York, pp. 3-10.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019), *BERT: Pre-training of deep bi-directional transformers for language understanding*, in Burstein, J., Doran, C., and Solorio, T. (2019, eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Vol. 1, Association for Computational Linguistics, Stroudsburg, PA, pp. 4171-4186.

Eger, S., Gleim, R. and Mehler, A. (2016), *Lemmatization and morphological tagging in German and Latin: A comparison and a survey of the state-of-the-art*, in Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, J., Moreno, A., Odijk, J. and Piperidis, S. (2016, eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, European Language Resources Association (ELRA), Paris, pp. 1507-1513.

Franzini, G., Peverelli, A., Ruffolo, P., Passarotti, M., Sanna, H., Signoroni, E., Ventura, V. and Zampedri, F. (2019), Nunc Est Aestimandum. *Towards an evaluation of the Latin WordNet*, in Bernardi, R., Navigli, R. and Semeraro, G. (2019, eds.), *CLiC-it 2019, Italian Conference on Computational Linguistics: Proceedings of the Sixth Italian Conference on Computational Linguistics: Bari, Italy, November 13-15, 2019*, RWTH Aachen, Aachen, § 33.

Gleim, R., Eger, S., Mehler, A., Uslu, T., Hemati, W., Lücking, A., Henlein, A., Kahlsdorf, S. and Hoenen, A. (2019), *A practitioner's view: A survey and comparison of lemmatization and morphological tagging in German and Latin*, in «Journal of Language Modeling», 7, 1, pp. 1-52.

Gleim, R., Mehler, A. and Ernst, A. (2012), *SOA implementation of the eHumanities Desktop*, in *Proceedings of the Workshop on Service-oriented Architectures (SOAs) for the Humanities: Solutions and Impacts*, Digital Humanities, London, pp. 24-29.

Halliday, M.A.K. and Hasan, R. (1976), *Cohesion in English*, Longman, London.

Haug, D.T. and Jøhndal, M. (2008), *Creating a parallel treebank of the old Indo-European bible translations*, in Sporleder, C. and Ribarov, K. (2008, eds.), *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, European Language Resources Association (ELRA), Paris, pp. 27-34.

Hemati, W., Uslu, T. and Mehler, A. (2016), *TextImager: A distributed UIMA-based system for NLP*, in Watanabe, H. (2016, ed.), *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, The COLING 2016 Organizing Committee, Osaka, pp. 59-63.

Jakobson, R. (1971), *Selected Writings II. Word and Language*, Mouton, The Hague.

Jiang, C., Yu, H.-F., Hsieh, C.-J. and Chang, K.-W. (2018), *Learning word embeddings for low-resource languages by PU learning*, in Walker, M., Ji, H. and Stent, A. (2018, eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Vol. 1: *Long Papers*, Association for Computational Linguistics, Stroudsburg, PA, pp. 1024-1034.

Jordan, M.D. (1995, ed.), *Patrologia Latina Database*, Chadwyck-Healey, Cambridge.

Joulin, A., Grave, E., Bojanowski, P. and Mikolov, T. (2017), *Bag of tricks for efficient text classification*, in Lapata, M., Blunsom, P. and Koller, A. (2017, eds.), *Proceedings of the 15th Conference of the EACL*. Vol. 2: *Short Papers*, Association for Computational Linguistics, Stroudsburg, PA, pp. 427-431.

Juršic, M., Mozetic, I., Erjavec, T. and Lavrac, N. (2010), LemmaGen*: Multilingual lemmatisation with induced ripple-down rules*, in «Journal of Universal Computer Science», 16, 9, pp. 1190-1214.

Jussen, B., Mehler, A. and Ernst, A. (2007), *A corpus management system for historical semantics*, in «Sprache und Datenverarbeitung. International Journal for Language Data Processing», 31, 1-2, pp. 81-89.

Jussen, B. (2009), *Perspektiven der Verwandtschaftsforschung fünfundzwanzig Jahre nach Jack Goodys »Entwicklung von Ehe und Familie in Europa«*, in Spiess, K.-H. (2009, *Hrsg.*), *Die Familie in der Gesellschaft des Mittelalters, Vorträge und Forschungen*, Thorbecke, Ostfildern, pp. 275-324.

Komninos, A. and Manandhar, S. (2016), *Dependency based embeddings for sentence classification tasks*, in Knight, K., Nenkova, A. and Rambow, O. (2016, eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, Stroudsburg, PA, pp. 1490-1500.

Kondratyuk, D., Gavenciak, T., Straka, M. and Hajic, J. (2018), LemmaTag*: Jointly tagging and lemmatizing for morphologically-rich languages with* BRNN*s*, in Riloff, E., Chiang, D., Hockenmaier, J. and Tsujii, J. (2018, eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Stroudsburg, PA, pp. 4921-4928.

Kondratyuk, D. and Straka, M. (2019), *75 languages, 1 model: Parsing universal dependencies universally*, in Inui, K., Jiang, J., Ng, V. and Wan, X. (2019, eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, Stroudsburg, PA, pp. 2779-2795.

Koster, C.H.A. and Verbruggen, E. (2002), *The* AGFL *Grammar Work Lab*, in Demetriou, C.G. (2002, ed.), *Proceedings of the* FREENIX *Track: 2002* USENIX *Annual Technical Conference*, USENIX Association, Berkeley, CA, pp. 13-18.

Levy, O. and Goldberg, Y. (2014), *Dependency-based word embeddings*, in Toutanova, T. and Wu, H. (2014, eds.), *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Vol. 2, Association for Computational Linguistics, Stroudsburg, PA, pp. 302-308.

Ling, W., Dyer, C., Black, A. and Trancoso, I. (2015), *Two/too simple adaptations of word2vec for syntax problems*, in Mihalcea, R., Chai, J. and Sarkar, A. (2015, eds.), *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, New York, pp. 1299-1305.

Mehler, A., Diewald, N., Waltinger, U., Gleim, R., Esch, D., Job, B., Küchelmann, T., Pustylnikov, O. and Blanchard, P. (2011), *Evolution of Romance language in written communication: Network analysis of late Latin and early Romance corpora*, in «Leonardo», 44, 3, pp. 244-245.

Mehler, A., Gleim, R., Hemati, W. and Uslu, T. (2017), *Skalenfreie online soziale Lexika am Beispiel von Wiktionary*, in Engelberg, S., Lobin, H., Steyer, K. and Wolfer, S. (2017, eds.), *Proceedings of 53rd Annual Conference of the Institut für Deutsche Sprache (*IDS*), March 14-16, Mannheim*, De Gruyter, Berlin, pp. 269-291.

Mehler, A., vor der Brück, T., Gleim, R. and Geelhaar, T. (2015), *Towards a network model of the coreness of texts: An experiment in classifying Latin texts using the* TTLab *Latin Tagger*, in Biemann, C. and Mehler, A. (2015, eds.), *Text Mining: From Ontology Learning to Automated Text Processing Applications, Theory and Applications of Natural Language Processing*, Springer, Berlin / New York, pp. 87-112.

Menge, H. (2009), *Lehrbuch der lateinischen Syntax und Semantik*, Wissenschaftliche Buchgesellschaft, Darmstadt.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013), *Efficient estimation of word representations in vector space* [preprint available online at *https://arxiv.org/abs/1301.3781*].

Mikolov, T., Yih, W. and Zweig, G. (2013), *Linguistic regularities in continuous space word representations*, in Vanderwende, L., Daumé, H., Kirchhoff, K. (2013, eds.), *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, Stroudsburg, PA, pp. 746-751.

Miller, G.A. (1995), *WordNet: A lexical database for English*, in «Communications of the ACM», 38, 11, pp. 39-41.

Miller, G.A. and Charles, W.G. (1991), *Contextual correlates of semantic similarity*, in «Language and Cognitive Processes», 6, 1, pp. 1-28.

Minozzi, S. (2017), *Latin WordNet, una rete di conoscenza semantica per il latino e alcune ipotesi di utilizzo nel campo dell'Information Retrieval*, in Mastrandrea, P. (2017, a cura di), *Strumenti digitali e collaborativi per le Scienze dell'Antichità*, Edizioni Ca' Foscari - Digital Publishing, Venezia, pp. 123-133.

Müller, T., Schmid, H. and Schütze, H. (2013), *Efficient higher-order* CRF*s for morphological tagging*, in Yarowsky, D., Baldwin, T., Korhonen, A., Livescu, K. and Bethard, S. (2013, eds.), *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Stroudsburg, PA, pp. 322-332.

Nivre, J., De Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R. and Zeman, D. (2016), *Universal Dependencies v1: A multilingual tree-bank collection*, in Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, J., Moreno, A., Odijk, J. and Piperidis, S. (2016, eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, European Language Resources Association (ELRA), Paris, pp. 1659-1666.

Passarotti, M. (2004), *Development and perspectives of the Latin morphological analyser* LemLat, in «Linguistica Computazionale», 20-21, pp. 397-414.

Passarotti, M. and Dell'Orletta, F. (2010), *Improvements in parsing the Index Thomisticus Treebank. Revision, combination and a feature model for Medieval Latin*, in Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M. and Tapias, D. (2010, eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, European Language Resources Association (ELRA), Paris, pp. 1964-1971.

Raible, W. (1981), *Von der Allgegenwart des Gegensinns (und einiger anderer Relationen), Strategien zur Einordnung semantischer Informationen*, in «Zeitschrift für romanische Philologie», 97, 1-2, pp. 1-40.

Rieger, B.B. (1989), *Unscharfe Semantik: die empirische Analyse, quantitative Beschreibung, formale Repräsentation und prozedurale Modellierung vager Wortbedeutungen in Texten*, Lang, Frankfurt am Main.

Rubenbauer, H., Hofmann, J.B. and Heine, R. (2009), *Lateinische Grammatik*, Oldenbourg, Bamberg.

Saussure, F. (1916), *Cours de linguistique générale*, edited by C. Bally and A. Sechehaye, Payot, Lausanne / Paris.

Schadt, H. (1982), *Die Darstellungen der Arbores consanguinitatis und der Arbores affinitatis: Bildschemata in juristischen Handschriften: Teilw. zugl.*, Tübingen, Univ., Diss., 1973, Wasmuth, Tübingen.

Schmid, H. (1994), *Probabilistic part-of-speech tagging using decision trees*, in Jones, D. and Somers, H. (1994, eds.), *New Methods in Language Processing Studies in Computational Linguistics*, UCL Press, London, pp. 154-164.

Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J. and Catanzaro, B. (2019), *Megatron-*LM*: Training multi-billion parameter language models using* GPU *model parallelism* [preprint available online at *https://arxiv.org/abs/1909.08053*].

Silva, A. and Amarathunga, C. (2019), *On learning word embeddings from linguistically augmented text corpora*, in Dobnik, S., Chatzikyriakidis, S. and Demberg, V. (2019, eds.), *Proceedings of the 13th International Conference on Computational Semantics - Short Papers*, Association for Computational Linguistics, Stroudsburg, PA, pp. 52-58.

Sowa, J.F. (2000), *Knowledge Representation: Logical, Philosophical, and Computational Foundations*, Brooks / Cole, Pacific Grove, CA.

Stoeckel, M., Henlein, A., Hemati, W. and Mehler, A. (2020), *Voting for PoS tagging of Latin texts: Using the flair of FLAIR to better Ensemble Classifiers by example of Latin*, in Liebeskind, C. and Liebeskind, S. (2020, eds.), *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages, Marseille, France*, European Language Resources Association (ELRA), Paris, pp. 130-135.

Straka, M. and Straková, J. (2017), *Tokenizing, PoS tagging, lemmatizing and parsing UD 2.0 with UDPipe*, in Hajič, J. and Zeman, D. (2017, eds.), *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Association for Computational Linguistics, Stroudsburg, PA, pp. 88-99.

Uslu, T., Mehler, A. and Baumartz, D. (2019), *Computing classifier-based embeddings with the help of text2ddc*, in *Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing, (CICLing 2019)* [available online at *https://www.researchgate.net/publication/332877162_Computing_Classifier-based_Embeddings_with_the_Help_of_text2ddc*].

Veremyev, A., Semenov, A., Pasiliao, E.L. and Boginski, V. (2019), *Graph-based exploration and clustering analysis of semantic spaces*, in «Applied Network Science», 4, 1, § 104.

vor der Brück, T. and Mehler, A. (2016), *TLT-CRF: A lexicon-supported morphological tagger for Latin based on conditional random fields*, in Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, J., Moreno, A., Odijk, J. and Piperidis, S. (2016, eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, European Language Resources Association (ELRA), Paris, pp. 1514-1519.

Yaskevich, A., Lisitsina, A., Katricheva, N., Zhordaniya, T., Kutuzov, A. and Kuzmenko, E. (2019), *Vec2graph: A Python library for visualizing word embeddings as graphs*, in van der Aalst, W., Batagelj, V., Ignatov, D.I., Khachay, M., Kuskova, V., Kutuzov, A., Kuznetsov, S.O., Lomazova, I.A., Loukachevitch, N., Napoli, A., Pardalos, P.M., Pelillo, M., Savchenko, A.V. and Tutubalina, E. (2019, eds.), *Analysis of Images, Social Networks and Texts. AIST 2019*, Springer, Cham.

Zapp, H. (1980), *Bann*, in *Lexikon des Mittelalters*. Vol. 1, Artemis Verlag, München / Zurich, pp. 1416-1417.

Zapp, H. (1989), *Exkommunikation*, in *Lexikon des Mittelalters*. Vol. 4, Artemis Verlag, München / Zurich, p. 170.

Zapp, H. (1991), *Interdikt*, in *Lexikon des Mittelalters*. Vol. 5, Artemis Verlag, München / Zurich, pp. 466-467.

Alexander Mehler
Department of Computer Science and Mathematics
Goethe Universität Frankfurt am Main
Robert-Mayer-Straße 10
D-60325 Frankfurt am Main
mehler@em.uni-frankfurt.de

Bernhard Jussen
Historical Seminar
Goethe Universität Frankfurt am Main
Norbert Wollheim Platz 1
60629 Frankfurt am Main
jussen@em.uni-frankfurt.de

Tim Geelhaar
Historical Seminar
Goethe Universität Frankfurt am Main
Norbert Wollheim Platz 1
60629 Frankfurt am Main
geelhaar@em.uni-frankfurt.de

Alexander Henlein
Department of Computer Science and Mathematics
Goethe Universität Frankfurt am Main
Robert-Mayer-Straße 10
D-60325 Frankfurt am Main
henlein@em.uni-frankfurt.de

Giuseppe Abrami
Department of Computer Science and Mathematics
Goethe Universität Frankfurt am Main
Robert-Mayer-Straße 10
D-60325 Frankfurt am Main
abrami@em.uni-frankfurt.de

Daniel Baumartz
Department of Computer Science and Mathematics
Goethe Universität Frankfurt am Main
Robert-Mayer-Straße 10
D-60325 Frankfurt am Main
baumartz@stud.uni-frankfurt.de

Tolga Uslu
Department of Computer Science and Mathematics
Goethe Universität Frankfurt am Main
Robert-Mayer-Straße 10
D-60325 Frankfurt am Main
uslu@em.uni-frankfurt.de

Wahed Hemati
Department of Computer Science and Mathematics
Goethe Universität Frankfurt am Main
Robert-Mayer-Straße 10
D-60325 Frankfurt am Main
hemati@em.uni-frankfurt.de

SSL

Ensemble lemmatization with the Classical Language Toolkit

Patrick J. Burns

Abstract

Because of the less-resourced nature of historical languages, non-standard solutions are often required for natural language processing tasks. This article introduces one such solution for historical-language lemmatization, that is the Ensemble lemmatizer for the Classical Language Toolkit, an open-source Python package that supports NLP research for historical languages. Ensemble lemmatization is the most recent development at *CLTK* in the repurposing and refactoring of an existing method designed for one task, specifically the backoff method as used for part-of-speech tagging, for use in a different task, namely lemmatization. This article argues for the benefits of ensemble lemmatization, specifically, flexible tool construction and the use of all available information to reach tagging decisions, and presents two use cases.

Keywords: lemmatization, natural language processing, Latin, Classical Language Toolkit.

1. *Introduction*

Because of the 'less-resourced' nature of historical languages, specifically due to what is often a paucity of extant text, limited availability of corpora and annotated data, as well as the incompatibility of tools that are available, non-standard solutions are often required for core natural language processing (NLP) tasks[1]. This article offers one such approach to the task of lemmatization, or «the process of transforming any word form into a corresponding, conventionally defined 'base' form» (Sprugnoli *et al.*, 2020: 105) developed for the Classical Language Toolkit (*CLTK*), an open-source Python package that supports NLP research for historical languages with text-analysis pipeline components, including lemmatizers

[1] For a definition of 'less-resourced' with reference to historical languages, see Piotrowski (2012: 85). In keeping with the theme of this special issue, this article will focus on Latin lemmatization, though the tools described here are under development, or can be adapted for use, for other languages.

Received: January 2020
Accepted: April 2020

(Johnson, 2020)[2]. I discuss here an 'ensemble' lemmatization method for Latin developed for *CLTK*, arguing for the benefits of this approach. By ensemble, I mean that the lemmatizer described is in fact a series of sub-lemmatizers that are deployed in unison with a selection mechanism included to limit the output to a single probable lemma or single group of probable lemmas. Following the example of combining results from more than one classifier in machine-learning setups, this version is called the Ensemble lemmatizer[3].

2. *Background*

Because of the lexicographical tradition for many historical languages, including Latin, lemmatization is of primary importance for NLP work on these languages; it is the 'fundamental annotation step' that allows related word forms, often forms with extensive morphological variation, to be grouped under a single identifier[4]. With respect to historical languages, Latin is well-served by off-the-shelf lemmatization tools, interfaces, web services, and desktop applications, including Collatinus, *LatMor*, *Lemlat*, Morpheus, and Whitaker's Words, among others; tools such as Stanza and TreeTagger can be also be included as language-independent tools that support Latin[5].

[2] For a description of text-analysis pipelines and components for historical languages, see Burns (2019). For related material on basic language resource kits, including material pertaining specifically to Latin, see Krauwer (2003); Passarotti (2010: 29); McGillivray (2014: 19-30). *CLTK* currently supports lemmatization for Ancient Greek, Latin, Old English and Old French; tool coverage for different *CLTK* languages can be found in the project's documentation: *http://docs.cltk.org/en/latest/*.

[3] See, for example, Dietterich (2000: 13): «Ensembles are well-established as a method for obtaining highly accurate classifiers by combining less accurate ones». For other examples of an ensemble approach used for Latin lemmatization and part-of-speech tagging, see Stoeckel *et al.* (2020) and Wu and Nicolai (2020).

[4] Mambrini and Passarotti (2019: 73); this article offers an excellent discussion of the lemma as an organizing principle for language tasks and the challenges therein. See Eger *et al.* (2015; 2016) and Gleim *et al.* (2019) for recent surveys of approaches to Latin lemmatization. Heslin (2019) contains a discussion of the challenges of automated Latin lemmatization in a literary critical context. Lemmatization, including specifically the disambiguation of homonymous word forms, has a significantly longer pre-computational tradition dating back to antiquity; see, for example, Dickey (2010: 193-201).

[5] Collatinus: Ouvrard (2010); *LatMor*: Springmann *et al.* (2016); *Lemlat*: Passarotti *et al.* (2017); Morpheus: Crane (1991), originally developed for Greek and later adapted for Latin; Words: Whitaker (1993); Stanza: Qi *et al.* (2020); TreeTagger: Schmid (1994). These lemmatizers are described in more detail in Burns (2019: 166-167).

Yet for the most part these lemmatizers are contextless taggers. That is, they provide lemma information based solely on the value of an isolated token, making no attempt to disambiguate returned tags using information such as the preceding or following words[6]. Accordingly, these tools can perform poorly on lemmatization tasks that would pose little challenge to a competent reader of Latin, as for example with the disambiguation of *ius* ("law") and *ius* ("broth, soup")[7].

Methods used in recent research on historical-language lemmatization include lexicon-assisted tagging and transformation rule induction, joint lemmatization and part-of-speech (PoS) tagging, as well as lemmatization as a neural-network-assisted string-transduction task[8]. With respect to the latter, research in historical-language lemmatization, following larger trends in NLP research generally, has taken a turn toward neural networks and deep learning approaches. These approaches, using either word- or character-level embeddings, often in conjunction with PoS tagging and dependency parsing, represent or near state-of-the-art performance for many languages[9]. Furthermore, neural-network approaches that take advantage of sentence-level context are proving to be especially effective, especially with respect to disambiguation (Bergmanis and Goldwater, 2018; Kestemont *et al.*, 2017; Manjavacas *et al.*, 2019)[10]. Another direction that has emerged in lemmatization for historical languages is their inclusion in recent large multilingual lemmatization studies due to their presence in the Universal Dependency treebanks (Nivre *et al.*, 2018)[11].

[6] See, for example, the notice in Passarotti *et al.* (2017: 25) on word form analysis using the *Lemlat* lemmatization tool: «Given an input word form that is recognised by *Lemlat*, the tool produces in output the corresponding lemma(s) [...] No contextual disambiguation is performed».

[7] It should be noted that intentional ambiguity is a nuance that lies outside the scope of computer-assisted approaches to lemmatization, at least as it is conceived of as an NLP task. For an overview of intentional ambiguity in Latin literature, see Fontaine *et al.* (2018), and pages xi-xii in particular on wordplay involving the ambiguity of *ius*.

[8] See, for example, Eger *et al.* (2015) and related work in Juršič *et al.* (2010), Bary *et al.* (2017), and Manjavacas *et al.* (2019), respectively.

[9] See, for example, Kondratyuk *et al.* (2018), Malaviya *et al.* (2019), Straka *et al.* (2019a), Straka and Straková (2020), and Celano (2020).

[10] See also, Chrupała (2006) on the usefulness of continuous text for the lemmatization of out-of-vocabulary words.

[11] Historical languages other than Latin, such as Ancient Greek, Coptic, Old French, and Old Church Slavonic, are also represented in version 2.3 of Universal Dependencies. For examples of recent multilingual shared task studies including Latin results, see Zeman *et al.* (2018) and Straka *et al.* (2019b).

In summary, despite advances, significant challenges still remain in historical-language lemmatization, in particular concerning the disambiguation of homonyms and the handling of unseen vocabulary, that is words that appear neither in a lexicon or in the training data used by the lemmatizer[12]. Moreover, there remains the question of whether 'lemma' is a stable enough category to be treated in a truly language-independent way and, for that reason, whether a lemmatizer should be designed to allow for a more flexible definition of the term[13]. Ensemble lemmatization works to address these challenges through flexibility of construction and the ability to combine results derived from a wide range of data sources, including lexicons, sentence-level training data, lists of regular expression patterns, and the output of other lemmatizers, among other sources[14].

3. *Lemmatizer construction with the Classical Language Toolkit*

Most approaches to historical-language lemmatization involve (i) taking an input, either a single token out of context or a token with its adjacent characters or words, (ii) performing a lookup of this token in a lexicon or otherwise analyzing this token, and (iii) returning a lemma or list of potential lemmas. Such approaches to lemmatization tend to share a certain fixity in design; that is, they tend to rely on a specific lexical data source or apply a specific set of rule-based transformations, and so on. Accordingly, the inter-

[12] Rosa and Žabokrtský (2019), for example, report 'deteriorations' on error reduction in unsupervised lemmatization for Latin. All the same, it is worth acknowledging how much progress has been made in this area since Ireland (1976: 46): «The present author knows of no system that as yet offers complete automatic lemmatization». If anything, this present author knows of several systems offering 'complete' automatic lemmatization; the focus of the current work is instead boosting accuracy, improving disambiguation, addressing a wider range of language domains, and handling the longest of long-tail vocabulary.

[13] See Knowles and Don (2004) on the difficulty of generalizing the idea of lemmatization across different languages, in particular English, Latin, Arabic, and Malay.

[14] The combination of multiple lemmatization strategies has something in common with the 'hybrid approach' described in Boudchiche and Mazroui (2019) which uses a two-pass lemmatization strategy: the first pass lemmatizes words out-of-context, a second pass uses a statistical method to disambiguate lemmas in context. Sychev and Penskoy (2019) describe a process for algorithmically «selecting different lemmatizers for different words» in English. For an early example of a staged approach to computer-assisted lemmatization, see Krause and Willée (1981). See also Romero (2019) for an example of 'modular design' in the construction of lemmatizers for Spanish and other languages.

nals of the lemmatization process are not exposed to the user[15]. The *CLTK*, on the other hand, offers options for lemmatization that specifically expose the lemmatizer construction process to the user, allowing for all intents and purposes an unlimited number of lexicons, rule definitions, or other tagging strategies to be combined and coordinated to reach a decision about the optimal choice of lemma for a given token.

3.1. *Backoff lemmatization*

Flexible lemmatizer construction was first introduced to the *CLTK* with the Backoff lemmatizer[16]. The main innovation of the Backoff lemmatizer was the repurposing of an existing method designed for one NLP task, specifically the backoff method as used for PoS tagging, for use in a different task, namely lemmatization[17]. In its original definition in the Natural Language Toolkit, sequential backoff tagging allows users to construct a PoS tagger from a set of sub-taggers (Bird *et al.*, 2015)[18]. A base tagger, called SequentialBackoffTagger, defines the backoff logic as follows: the first sub-tagger in the sequence attempts to tag a given token and, if it is unable to do so, the next sub-tagger in the sequence (that is, the 'backoff' tagger) is tried and so on, until either a token is successfully tagged or the sequence ends. Various sub-taggers make use of different tagging strategies, including the use of frequency data from annotated sentences, custom lexicons, or lists of regular expressions patterns, among other resources, to assign tags. The effectiveness of Sequential Backoff Tagger resides not in any specific sub-tagger but in their combined deployment, since subsequent taggers compensate for the gaps in coverage of previous ones.

[15] It is true that most of the available tools offer some degree of customization with respect to the lemmatization process, even if they lack the flexibility of construction and choice of parameters offered by the *CLTK* lemmatizers. For example, *Lemlat* and Collatinus have parameters available for choosing the lexical basis for analyzing tokens; see *https://github.com/CIRCSE/LEMLAT3/wiki/2.-Use* and *https://outils.biblissima.fr/en/collatinus-web/* respectively.

[16] The Backoff lemmatizer for Latin was developed as part of a 2016 Google Summer of Code project; see the project description here: *https://summerofcode.withgoogle.com/archive/2016/projects/6499722319626240*. The source code can be found in the Lemmatize module at *https://github.com/cltk/cltk/tree/master/cltk/lemmatize*.

[17] The basic design of the Backoff lemmatizer is given in Burns (2016) with additional description in the section 'Lemmatization as reading' in Burns *et al.* (2019). The discussion here of the Backoff lemmatizer is meant to provide context for understanding the motivation for the development of the Ensemble lemmatizer.

[18] The source code for SequentialBackoffTagger and its subclasses can be found at *https://www.nltk.org/_modules/nltk/tag/sequential.html*; see also Perkins (2014: 92-93).

The repurposing of SequentialBackoffTagger for lemmatization makes sense because at its heart lemmatization is a tagging task (Gesmundo and Samardžić, 2012). That said, as opposed to the well-bounded task of PoS tagging, lemmatization is an infinite tagging task. There are 17 tags in the Universal PoS tagset and 36 in the Penn Treebank PoS tagset[19]. Even with largely fixed-corpus languages such as Ancient Greek and Latin, there are a nearly infinite number of word forms that could be mapped to a lemma, something made clear, for example, by the hundreds of 'new' words published in the supplements to Liddell and Scott's *Greek-English Lexicon*[20]. Accordingly, from a tagging perspective, the performance of a lexicon-based approach can only be improved by expanding lexicon coverage, and even at that, the direction and degree of this expansion would be difficult, if not impossible, to predict. So, for example, the Latin coinage *telecommunicationis* ("of telecommunication") as found in the Latin Wikipedia article about the telephone will not be tagged by any off-the-shelf Latin lemmatizer[21]. Still, this word would likely be lemmatized correctly if a regular-expression-based lemmatizer is included in the backoff chain, since its genitive singular word ending (*-ationis*) can be mapped predictably to the nominative singular form that is traditionally used for reporting Latin noun lemmas[22]. It is this combination of data-driven and rules-based strategies that makes backoff tagging an effective approach to lemmatization.

That said, backoff tagging has a major disadvantage. SequentialBackoffTagger takes a binary approach to tagging; that is, at any given point in the backoff chain, a tagger either assigns a tag or it does not. If a tag is assigned, the sequence is terminated and the tagger moves onto the next token. Foreshortening the backoff chain in this way improves processing speed, but at the cost of loss of information from the unused taggers. Moreover, the arrangement of sub-lemmatizers in the backoff chain can have a hard to predict effect on the results.

[19] For the Universal PoS tagset, see *https://universaldependencies.org/u/pos/*; for the Penn tagset, see Santorini (1995).

[20] See Glare and Thompson (1996), itself a revision of an earlier version from 1968. Eger *et al.* (2016: 1507 n. 2) also notes that the lexicons «cannot store an infinite number of words».

[21] See *https://la.wikipedia.org/wiki/Telephonum*: *Telephonum* [...] *est instrumentum telecommunicationis quo homines per longa spatia inter se loqui possunt* "The telephone is an instrument of telecommunication with which people are able to speak to each other over long distances".

[22] See Diederich (1939: 21-30) for a statistical evaluation of the use of Latin word endings to determine lemmas.

3.2. *Ensemble lemmatization*

In order to avoid the loss of potentially useful information from sub-lemmatizers further down the backoff chain, the Backoff lemmatizer has been refactored so as not to terminate upon the first successful tagging. The resulting tool is the Ensemble lemmatizer[23]. With this setup, all tokens are tagged by all sub-lemmatizers. No tagging information is lost. At the completion of the tagging operation, a list of potential lemmas is returned, and, if requested, a selection mechanism can be used to limit this output to a single probable lemma.

The advantage of complete multiple-pass tagging is that all available information provided by sub-lemmatizers in the sequence is retained and, as such, can be used to make a final determination. Here is a simple example based on Cicero's *De domo suo* 39: *Infirmas igitur tu acta C. Caesaris?*, "Are you therefore weakening Gaius Caesar's decrees?".

We can construct an Ensemble lemmatizer using two sub-lemmatizers, namely a lexicon-based lemmatizer (EnsembleDictLemmatizer) with a lexicon mapping the token *infirmas* to the lemma *infirmus* and regular-expression-based lemmatizer (EnsembleRegexpLemmatizer) with a pattern that replaces tokens ending in *-as* (and other present active endings for first conjugation Latin verbs), in that order (reading from the bottom up):

(1) regexp_ensemble_lemmatizer = EnsembleRegexpLemmatizer(patterns= [('(.)a(s|t|mus|tis|nt)$', '\1o')], backoff = None)

dict_ensemble_lemmatizer = EnsembleDictLemmatizer(dictionary = {'infirmas': 'infirmus'}, backoff = regexp_ensemble_lemmatizer)

As opposed to the backoff setup, the fact that the lexicon-based lemmatizer tags *infirmas* (incorrectly) as a form of the adjective *infirmus* ("weak") on the first pass does not prevent it from also tagging the token (correctly) as the verb *infirmo* ("to weaken") on the second pass. Some selection mechanism needs to be used to perform the disambiguation, whether frequency distributions from training data, probabilities assigned to word-ending patterns, contextual semantics, confidence scores based on dependency parsing[24], and so on. Again, this is a trivial example designed to explain how the

[23] The source code can be found in the Lemmatize module at *https://github.com/cltk/cltk/tree/master/cltk/lemmatize*.

[24] This sentence provides an excellent example of how dependency tree information could be combined with traditional approaches to reading Latin to assist with lemma disambiguation as

Ensemble lemmatizer works and in particular how it works differently than the Backoff lemmatizer. An example showing the clear advantage of the ensemble setup is offered in Section 4.

A final point on the Ensemble lemmatizer. While the example above shows only two main types of sub-lemmatizers, that is lexicon-based and regular-expression-based lemmatizers, the 'building block'-style design of this lemmatizer allows for the development of any number of sub-lemmatizers. By subclassing SequentialBackoffTagger and overriding the 'tag' method with a different method of determining a lemma from a token, any lemmatization algorithm can be incorporated into the Ensemble lemmatizer. As long as a subclass of one of the lemmatizers (i) accepts a list of tokens as its input and (ii) provides a list of lemmas as its output, it can be added to the lemmatization chains.

A specific kind of sub-lemmatizer is ideal for development under this 'building block' logic, namely wrappers, that is classes or functions that allow external code to be used locally, written for existing lemmatization tools[25]. As noted above, there are several off-the-shelf options available for lemmatizing Latin texts, but at present their results cannot be effectively collated and evaluated without some sort of ad hoc post-processing. Moreover, these tools can be incompatible with each other or otherwise not customizable or extensible[26]. This is because each tool is envisioned as a self-sufficient solution for the task. Ensemble lemmatization reconceives them as part of a coordinated lemmatization solution, the combined results of which can be easily and directly incorporated into a tagging workflow. So, rather than having to choose TreeTagger or *Lemlat*, wrapper-based sub-lemmatizers can be chained together so that both are used, leveraging the strengths of each[27].

infirmas ("you weaken") is the only eligible verb in this sentence, not to mention that the explicit (and unambiguous) subject *tu* ("you") confirms the requirement of a second-person singular verb in the sentence. Ensemble lemmatizer development following these kinds of traditional reading approaches is ongoing; see McCaffrey (2006), for example, on disambiguation in reading Latin as well as the discussion of 'philological method' in Section 5 below.

[25] For a general discussion of wrappers in the *CLTK* pipeline, see Burns (2019: 171-172). On wrappers as a best practice when working with third-party software, see Martin (2009: 109).

[26] Addressing interoperability is a primary objective of the Linking Latin (*LiLa*) project; see the *LiLa* objectives here *https://lila-erc.eu/about* as well as in Mambrini and Passarotti (2019).

[27] An example of a chained-together wrappers is given below in Section 4.

4. *Use cases*

While high accuracy is obviously a goal of any NLP tool, the more important contribution of ensemble lemmatization comes with the coordination of results made possible by its modular, flexible construction which allows for a greater degree of customization depending on the language being processed (and the availability of supporting resources for this language) as well as the domain being studied, the research question under consideration, and so on[28]. To illustrate the benefits of this coordination, modularity, and flexibility, I offer two uses cases: (i) the lemmatization of a text likely to pose a significant challenge to existing tools, namely a Latin translation of Lewis Carroll's 'Jabberwocky' and (ii) the use of the Ensemble lemmatizer to combine effectively existing tools.

4.1. *Lemmatizing 'Jabberwocky' with the Ensemble lemmatizer*

The handling of unseen vocabulary is a challenge for lemmatizers. For historical languages, this challenge is particularly acute because, not only are they often less-resourced in general, but their resources can be especially limited for variations of dialect, period, and so on[29]. The example here illustrates this with an extreme case, namely lemmatizing a Latin translation of Lewis Carroll's nonsense poem, 'Jabberwocky', by C. H. Carruthers[30]. Here are the opening lines: *Est brilgum: tovi slimici / in vabo tererotitant* "'Twas brillig, and the slithy toves / did gyre and gimble in the wabe". Some words here would present no difficulty to any Latin lemmatizer: *est* and *in*. The remaining words however will understandably not appear in any Latin lexicon and for this reason off-the-shelf solutions will be unlikely to yield results. At the same time, a competent reader of Latin can lemmatize this text with minimal difficulty through additional interpretative strategies. *Tererotitant* can only be lemmatized as *tererotito*; the Latin reader knows this because the -(*t*)*ant* ending, a marker of the third-person active plural, can be meaningfully transformed to

[28] For a look into the current state of evaluation for Latin lemmatization methods, see Sprugnoli *et al.* (2020) and the participating papers in the *EvaLatin* 2020 campaign.

[29] See Kestemont and De Gussem (2017) for using a neural-network approach to handle historical-language variation, and in particular, Medieval Latin orthography.

[30] This translation appears as *Jabberwocky: An Alternative Version* in Carroll (1966: 132-133). For background on these translations and others, see Imholtz (1987); Van Dam (1982). For an example of NLP methods used on this poem, see Feldman (1999).

the first-person present indicative active forms traditionally used for verb lemmas[31]. Accordingly, if we set up a backoff sequence that reflects the processes of a competent reader, we can make meaningful inroads in lemmatizing this text:

(2) regexp_lemmatizer = EnsembleRegexpLemmatizer(patterns = [('(.)a(s|t|mus|tis|nt)$', '\1o')], backoff = None)
dict_lemmatizer = EnsembleDictLemmatizer(dictionary = {'est': 'sum', 'in': 'in'}, backoff = regexp_lemmatizer)

Additional patterns could be written for other nonsense words in the poem: *vorpalem* to *vorpalis*, *Unguimanu* to *Unguimanus*, *gaudiferum* to *gaudifer*, *praehilare* to *praehilaris*, and so on[32]. Admittedly, the lemmatization of Latin nonsense poetry is a low-priority problem. Nevertheless, the issues raised by this problem, most especially dealing with unknown word forms and transforming them in a consistent, philologically sound manner, will surface whenever NLP tools are used on 'underserved domains' and an ensemble approach is well-equipped to handle this situation[33].

4.2. *Combining lexicons with the Ensemble lemmatizer*

As noted above, off-the-shelf Latin lemmatizers are generally envisioned as self-sufficient solutions for the task and as a result there is often no direct way to combine efficiently and aggregate the results of multiple tools. The Ensemble lemmatizer using wrappers written for existing tools can solve this problem. Here is an example based on the beginning of Book 12 of Ovid's *Metamorphoses*: *Nescius adsumptis Priamus pater Aesacon alis / vivere lugebat* "Father Priam was mourning for Aesacus, not realizing that he had assumed wings and was alive". If we set up a backoff chain with wrappers for Latin lemmatizers mentioned in Section 2 as follows:

[31] Sequence-modeling approaches could also be used to address this, though there would perhaps be a concern of adding unbounded noise to the textual noise inherent in nonsense poetry. See Kestemont and De Gussem (2017) for a discussion of 'computational hypercorrection' and the generation of 'unrecognisable form[s]'. Using a list of regular-expression-based replacement patterns that reflect traditional expectations about the morphological information found in word endings goes some way in mitigating this concern; see below on 'philological method' in Section 5. Moreover, a sequence-modeling-based wrapper could always be written for the Ensemble lemmatizer and could substitute for (or complement) the regular-expression-based lemmatizer in this sequence.

[32] A starter set of regular-expression-based replacement patterns for Latin can be found at *https://github.com/cltk/cltk/blob/master/cltk/lemmatize/latin/latin.py*.

[33] See Bamman (2017) on NLP for 'underserved domains'.

(3) lemlat = LemlatLemmatizer(backoff = None)
collatinus = CollatinusLemmatizer(backoff = lemlat)
words = WordsLemmatizer(backoff = collatinus)
morpheus = MorpheusLemmatizer(backoff = words)
latmor = LatmorLemmatizer(backoff = morpheus)
treetagger = TreeTaggerLemmatizer(backoff = latmor)

we get a better sense of how the coverage of each tool complements the others. Several words in this example pose no problem for any of the lemmatizers: *nescius*, *pater*, *vivere*, and *lugebat* are all tagged correctly and unambiguously as *nescius*, *pater*, *vivo*, and *lugeo*, respectively[34]. In other cases, an individual tagger fails to return a lemma, but this gap is covered by one of the other taggers: for example, TreeTagger does not return a lemma for *Priamus*, but Collatinus, *LatMor*, *Lemlat*, Morpheus, and Whitaker's Words all return the correct lemma. A token like *ne* (*Met*. 12.590) presents the opposite problem, as the tools return different sets of lemmas: *ne* (TreeTagger); *ne*, *neo* (Collatinus, *Lemlat*, Morpheus, Words); and *ne*, *nere* (*LatMor*). In this case, even a simple count-based vote would return the correct lemma *ne*, present six times across the results of the six lemmatizers[35]. Still, the more important point here is that the Ensemble lemmatizer provides a direct way of combining the output of multiple taggers and maximizing the amount of information available for determining the best choice.

5. *Conclusion*

5.1. *Ensemble lemmatization as a philological method*

As described above, the Ensemble lemmatizer offers technical advantages to the lemmatization of historical-language text. It is worth noting that this approach to lemmatization offers a theoretical advantage as well to the

[34] *LatMor* with its default settings tags *vivere* not as *vivo* but as the present active infinitive *vivere*. In testing this configuration, the *LatMor* wrapper normalized the output of verbs by re-lemmatizing these infinitives with another tool (here, namely, Collatinus). The normalization that can be built into the Ensemble wrappers can be seen as another benefit of the approach.

[35] Other options exist for resolving similar lists of possible lemmas. Gawley (2019), for example, presents a disambiguation method based on corpus frequencies and Heslin (2019) proposes a novel method for disambiguation that compares the lengths of lexicon entries for respective forms.

primary audience for *CLTK*'s tools, namely historical-language researchers, instructors, and students. I have argued before that backoff lemmatization «can be described as following a philological method» because it reflects the decoding strategies of the philologically trained reader of historical texts (Burns, 2018)[36]. That said, ensemble lemmatization demonstrates this even more clearly since it draws on multiple sources of information and makes use of all of them in arriving at a decision. This reflects, for example, the process of the textual critic who, through both a comprehensive accounting of word use in context and the relative frequency of tokens and their endings, is able to make philologically informed decisions about possible readings[37]. This also reflects the process of a Latin translator for whom working through a text with multiple passes can be an effective decipherment strategy, as one Latinist recommends: «Once you know what all the words can mean, re-read the Latin to [...] clarify what the words in the sentence [...] mean» (Hoyos, 2008). Yet another group of Latin teachers emphasize this progressive clarification as a «dynamic process which involves continual reconsideration of previous decisions and expectations», not unlike the process whereby the Ensemble lemmatizer accumulates potential lemmas before arriving at a decision about the most probable lemma or lemmas (Markus and Ross, 2004: 88). The backoff and the ensemble approaches to lemmatization, and the ensemble approach in particular, reflect established disciplinary practices for disambiguating words and acknowledge that this process often requires coordinated methods.

5.2. *Future directions*

The Ensemble lemmatizer discussed here is available at present for Latin, but is included in the 'Multilingual' section of the *CLTK* documentation, since the sub-lemmatizers can be used with any language for which supporting resources such as token-lemma lexicons, annotated

[36] For a discussion of decoding strategies as applicable to the study of Latin, see McCaffrey (2006; 2009); Russell (2018); see also Burns *et al.* (2019) on the relationship between lemmatization, literacy, and 'classical-language reading patterns'. A reviewer astutely points out that similar decoding strategies may be typical of language users generally in negotiating the meaning of words in context; I limit the discussion here to observations that have been made on this point concerning philological activities such as textual criticism and historical-language pedagogy.

[37] See Tarrant (2016: 57): «When choosing between or among equally well-attested variants, the editor may have recourse to a variety of potentially relevant factors». For an example of a systematic study of word endings in the context of textual criticism, see Håkanson (1982).

sentences, or regular expression patterns can be provided. A good next step would be the development of default sequences for the full range of languages covered by *CLTK* for which lemmatization is a core task. Another good step would be the development of more wrappers that can be used with the Ensemble lemmatizer, not only for off-the-shelf tools as discussed above, but also for the state-of-the-art methods discussed in Section 2. In the spirit of the 'all available information' approach of the Ensemble lemmatizer, it is not hard to see the benefit to *CLTK* users in being able to include these methods in backoff sequences and combine them with other methods.

Acknowledgments

Earlier versions of this article were presented at the First *LiLa* Workshop: Linguistic Resources & NLP Tools for Latin at Università Cattolica del Sacro Cuore in the summer of 2019 as well as at DH2018 in Mexico City and the Digital Classicist London Seminar in the summer of 2018. The author thanks the organizers, participants, and attendees of these events for their feedback. The author would also like to acknowledge the Classical Language Toolkit open-source development community, the Institute for the Study of the Ancient World Library, and the Quantitative Criticism Lab for their support. Lastly, the author thanks the anonymous reviewers of this article for their suggestions.

References

Bamman, D. (2017), *Natural Language Processing for the long tail*, in Lewis, R., Raynor, C., Forest, D., Sinatra, M. and Sinclair, S. (2017, eds.), *Digital Humanities 2017, DH 2017, Conference Abstracts, McGill University & Université de Montréal, Montréal, Canada, August 8-11, 2017*, Alliance of Digital Humanities Organizations (ADHO).

Bary, C., Berck, P. and Hendrickx, I. (2017), *A memory-based lemmatizer for Ancient Greek*, in *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage - DATeCH2017*, Association for Computing Machinery, New York, pp. 91-95.

Bergmanis, T. and Goldwater, S. (2018), *Context sensitive neural lemmatization with Lematus*, in Walker, M., Ji, H. and Stent, A. (2018, eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Asso-*

ciation for Computational Linguistics: Human Language Technologies. Vol. 1: *Long Papers*, Association for Computational Linguistics, Stroudsburg, PA, pp. 1391-1400.

Bird, S., Klein, E. and Loper, E. (2015), *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit* [available online at *https://www.nltk.org/book*, accessed on 21.05.2020].

Boudchiche, M. and Mazroui, A. (2019), *A hybrid approach for Arabic lemmatization*, in «International Journal of Speech Technology», 22, pp. 563-573.

Burns, P.J. (2016), *Wrapping up Google Summer of Code* [available online at *https://disiectamembra.wordpress.com/2016/08/23/wrapping-up-google-summer-of-code/*, accessed on 21.05.2020].

Burns, P.J. (2018), *Backoff lemmatization as a philological method*, in Girón Palau, J. and Russell, I.G. (2018, eds.), *Digital Humanities 2018, DH 2018, Book of Abstracts, El Colegio de México, UNAM, and RedHD, Mexico City, Mexico, June 26-29, 2018*, Red de Humanidades Digitales.

Burns, P.J. (2019), *Building a text analysis pipeline for Classical languages*, in Berti, M. (2019, ed.), *Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution*, De Gruyter, Berlin, pp. 159-176.

Burns, P.J., Hollis, L. and Johnson, K.P. (2019), *The future of ancient literacy: Classical Language Toolkit and Google Summer of Code*, in «Classics@», 17.

Carroll, L. (1966), *Ludovici Carroll fabella lepida in qua aliud Aliciae sominium narravit: Aliciae per speculum transitus (quaeque ibi invenit)*, Macmillan, London.

Celano, G.G.A. (2020), *A gradient boosting-Seq2Seq system for Latin PoS tagging and lemmatization*, in Sprugnoli, R. and Passarotti, M. (2020, eds.), *Proceedings of 1st Workshop on Language Technologies for Historical and Ancient Languages*, European Language Resources Association (ELRA), Paris, pp. 119-123.

Celano, G.G.A. *et al.* (2014), *The Ancient Greek and Latin Dependency Treebank v.2.1* [available online at *https://perseusdl.github.io/treebank_data*, accessed on 21.05.2020].

Chrupała, G. (2006), *Simple data-driven context-sensitive lemmatization*, in «Procesamiento del Lenguaje Natural», 37, pp. 121-127.

Crane, G. (1991), *Generating and parsing Classical Greek*, in «Literary and Linguistic Computing», 6, pp. 243-245.

Dickey, E. (2010), *The creation of Latin teaching materials in antiquity: A re-interpretation of P.sorb. Inv. 2069*, in «Zeitschrift für Papyrologie und Epigraphik», 175, pp. 188-208.

Diederich, P.B. (1939), *The Frequency of Latin Words and Their Endings*, The University of Chicago Press, Chicago.

Dietterich, T.G. (2000), *Ensemble methods in machine learning*, in Kittler, J. and Roli, F. (2000, eds.), *Multiple Classifier Systems. Proceedings of the First International Workshop, MCS 2000 (Cagliari, Italy, June 21-23, 2000)*, Springer, Berlin, pp. 1-15.

Eger, S., Gleim, R. and Mehler, A. (2016), *Lemmatization and morphological tagging in German and Latin: A comparison and a survey of the state-of-the-art*, in Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, J., Moreno, A., Odijk, J. and Piperidis, S. (2016, eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, European Language Resources Association (ELRA), Paris, pp. 1507-1513.

Eger, S., vor der Brück, T. and Mehler, A. (2015), *Lexicon-assisted tagging and lemmatization in Latin: A comparison of six taggers and two lemmatization methods*, in Zervanou, K., van Erp, M. and Alex, B. (2015, eds.), *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, Association for Computational Linguistics, Stroudsburg, PA, pp. 105-113.

Feldman, S. (1999), *NLP meets the Jabberwocky: Natural Language Processing in information retrieval*, in «ONLINE», 23.

Fontaine, M., McNamara, C.J. and Short, W.M. (2018), *Introduction*, in Fontaine, M., McNamara, C.J. and Short, W.M. (2018, eds.), Quasi labor intus*: Ambiguity in Latin Literature: Papers in Honor of Reginald Thomas Foster, OCD*, Paideia Institute for Humanistic Study, Middletown, DE, pp. ix-xxxi.

Gawley, J.O. (2019), *An unsupervised lemmatization model for Classical languages* [available online at *https://dev.clariah.nl/files/dh2019/boa/1007.html*, accessed on 21.05.2020].

Gesmundo, A. and Samardžić, T. (2012), *Lemmatisation as a tagging task*, in Li, H., Lin, C-Y., Osborne, M., Geunbae Lee, G. and Park, J.C. (2012, eds.), *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*. Vol. 1: *Short Papers*, Association for Computational Linguistics, Stroudsburg, PA, pp. 368-372.

Glare, P.G.W. and Thompson, A.A. (1996, eds.), *Greek-English Lexicon. Revised Supplement*, Clarendon Press, Oxford.

Gleim, R., Eger, S., Mehler, A., Uslu, T., Hemati, W., Lücking, A., Henlein, A., Kahlsdorf, S. and Hoenen, A. (2019), *A practitioner's view: A survey and comparison of lemmatization and morphological tagging in German and Latin*, in «Journal of Language Modelling», 7, pp. 1-52.

Håkanson, L. (1982), Homoeoteleuton *in Latin dactylic poetry*, in «Harvard Studies in Classical Philology», 86, pp. 87-115.

Heslin, P. (2019), *Lemmatizing Latin and quantifying the Achilleid*, in Coffee, N., Forstall, C., Milić, L.G. and Nelis, D. (2019, eds.), *Intertextuality in Flavian Epic Poetry*, De Gruyter, Berlin, pp. 389-408.

Hoyos, D. (2008), *The ten basic rules for reading Latin* [available online at *http://www.latinteach.com/Site/ARTICLES/Entries/2008/10/15_Dexter_Hoyos_-_The_Ten_Basic_Rules_for_Reading_Latin_files/Reading%20%26Translating%20Rules.pdf*, accessed on 21.05.2020].

Imholtz, A.A. (1987), *Latin and Greek versions of «Jabberwocky». Exercises in laughing and grief*, in «Rocky Mountain Review of Language and Literature», 41, pp. 211-228.

Ireland, S. (1976), *The computer and its role in classical research*, in «Greece & Rome», 23, pp. 40-54.

Johnson, K.P. (2020), *CLTK: The Classical Language Toolkit* [available online at *https://github.com/cltk/cltk*, accessed on 21.05.2020].

Juršič, M., Mozetič, I., Erjavec, T. and Lavrač, N. (2010), *LemmaGen: Multilingual lemmatisation with induced ripple-down rules*, in «Journal of Universal Computer Science», 16, pp. 1190-1214.

Kestemont, M. and De Gussem, J. (2017), *Integrated sequence tagging for medieval Latin using deep representation learning*, in Büchler, M. and Mellerin, L. (2017, eds.), *Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages*, «Journal of Data Mining & Digital Humanities» (special issue).

Kestemont, M., de Pauw, G., van Nie, R. and Daelemans, W. (2017), *Lemmatization for variation-rich languages using deep learning*, in «Digital Scholarship in the Humanities», 32, pp. 797-815.

Knowles, G. and Don, Z.M. (2004), *The notion of a "lemma": Headwords, roots and lexical sets*, in «International Journal of Corpus Linguistics», 9, pp. 69-81.

KONDRATYUK, D., GAVENČIAK, T., STRAKA, M. and HAJIČ, J. (2018), *LemmaTag: Jointly tagging and lemmatizing for morphologically-rich languages with BRNNs* [preprint available online at *https://arxiv.org/abs/1808.03703*].

KRAUSE, W. and WILLÉE, G. (1981), *Lemmatizing German newspaper texts with the aid of an algorithm*, in «Computers and the Humanities», 15, pp. 101-113.

KRAUWER, S. (2003), *The Basic Language Resource Kit (BLARK) as the first milestone for the Language Resources Roadmap*, in *Proceedings of the International Workshop Speech and Computer, Moscow, Russia*, pp. 8-15.

MALAVIYA, C., WU, S. and COTTERELL, R. (2019), *A simple joint model for improved contextual neural lemmatization* [preprint available online at *https://arxiv.org/abs/1904.02306*].

MAMBRINI, F. and PASSAROTTI, M. (2019), *Harmonizing different lemmatization strategies for building a knowledge base of linguistic resources for Latin*, in FRIEDRICH, A., ZEYREK, D., and HOEK, J. (2019, eds.), *Proceedings of the 13th Linguistic Annotation Workshop*, Association for Computational Linguistics, Stroudsburg, PA, pp. 71-80.

MANJAVACAS, E., KÁDÁR, A. and KESTEMONT, M. (2019), *Improving lemmatization of non-standard languages with joint learning* [preprint available online at *https://arxiv.org/abs/1903.06939*].

MARKUS, D.D. and ROSS, D.P. (2004), *Reading proficiency in Latin through expectations and visualization*, in «Classical World», 98, pp. 79-93.

MARTIN, R.C. (2009), *Clean Code: A Handbook of Agile Software Craftsmanship*, Prentice Hall, Upper Saddle River, NJ.

MCCAFFREY, D. (2006), *Reading Latin efficiently and the need for cognitive strategies*, in GRUBER-MILLER, J. (2006, ed.), *When Dead Tongues Speak: Teaching Beginning Greek and Latin*, Oxford University Press, New York, pp. 113-133.

MCCAFFREY, D. (2009), *When reading Latin, read as the Romans did*, in «Classical Outlook», 86, pp. 62-66.

MCGILLIVRAY, B. (2014), *Methods in Latin Computational Linguistics*, Brill, Leiden.

NIVRE, J., ABRAMS, M. and AGIĆ, Ž. (2018), *Universal Dependencies v.2.3* [available online at *http://hdl.handle.net/11234/1-2895*, accessed on 21.05.2020].

OUVRARD, Y. (2010), *Collatinus, lemmatiseur et analyseur morphologique de la langue latine*, in «ÉLA. Études de linguistique appliquée», 158, pp. 223-230.

Passarotti, M. (2010), *Leaving behind the less-resourced status. The case of Latin through the experience of the Index Thomisticus Treebank*, in Sarasola, K., Tyers, F.M. and Forcada, M.L. (2010, eds.), *Proceedings of the 7th SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-Resourced Languages, LREC 2010*, European Language Resources Association, La Valletta, pp. 27-32.

Passarotti, M., Budassi, M., Litta, E. and Ruffolo, P. (2017), *The* Lemlat *3.0 package for morphological analysis of Latin*, in Bouma, G. and Adesam, Y. (2017, eds.), *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, Linköping University Electronic Press, Linköping, pp. 24-31.

Perkins, J. (2014), *Python 3 Text Processing with NLTK 3 Cookbook*, Packt Publishing Ltd., Birmingham, U.K.

Piotrowski, M. (2012), *Natural Language Processing for Historical Texts*, Morgan & Claypool Publishers, San Rafael, CA.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J. and Manning, C.D. (2020), *Stanza: A Python natural language processing toolkit for many human languages* [preprint available online at *https://arxiv.org/abs/2003.07082*].

Romero, G. (2019), *Rethinking rule-based lemmatization* [available online at *https://www.youtube.com/watch?v=88zcQODyuko*, accessed on 21.05.2020].

Rosa, R. and Žabokrtský, Z. (2019), *Unsupervised lemmatization as embeddings-based word clustering* [preprint available online at *https://arxiv.org/abs/1908.08528*].

Russell, K. (2018), *Read like a Roman: Teaching students to read in Latin word order*, in «Journal of Classics Teaching», 19, pp. 17-29.

Santorini, B. (1995), *Part-of-speech tagging guidelines for the Penn Treebank Project* (3rd revision, 2nd printing) [available online at *https://catalog.ldc.upenn.edu/docs/LDC99T42/tagguid1.pdf*, accessed on 14.07.2020].

Schmid, H. (1994), *Probabilistic part-of-speech tagging using decision trees*, in *Proceedings of International Conference on New Methods in Language Processing, Manchester, UK*, pp. 44-49.

Springmann, U., Schmid, H. and Najock, D. (2016), *LatMor: A Latin finite-state morphology encoding vowel quantity*, in «Open Linguistics», 2, pp. 386-392.

Sprugnoli, R., Passarotti, M., Cecchini, F.M. and Pellegrini, M. (2020), *Overview of the EvaLatin 2020 evaluation campaign*, in Sprugnoli, R. and Passarotti, M. (2020, eds.), *Proceedings of 1st Workshop on Language Technologies for Historical and Ancient Languages*, European Language Resources Association (ELRA), Paris, pp. 105-110.

Stoeckel, M., Henlein, A., Hemati, W. and Mehler, A. (2020), *Voting for PoS tagging of Latin texts: Using the flair of FLAIR to better ensemble classifiers by example of Latin*, in Sprugnoli, R. and Passarotti, M. (2020, eds.), *Proceedings of 1st Workshop on Language Technologies for Historical and Ancient Languages*, European Language Resources Association (ELRA), Paris, pp. 130-135.

Straka, M., Straková, J. and Hajič, J. (2019a), *Czech text processing with contextual embeddings: PoS tagging, lemmatization, parsing and NER*, in Ekštein, K. (2019, ed.), *Text, Speech, and Dialogue. TSD 2019* (Lecture Notes in Computer Science, vol. 11697), Springer, Cham, pp. 137-150.

Straka, M., Straková, J. and Hajič, J. (2019b), *Evaluating contextualized embeddings on 54 languages in PoS tagging, lemmatization and dependency parsing* [preprint available online at *https://arxiv.org/abs/1908.07448*].

Straka, M. and Strakova, J. (2020), *UDPipe at EvaLatin 2020: Contextualized embeddings and treebank embeddings*, in Sprugnoli, R. and Passarotti, M. (2020, eds.), *Proceedings of 1st Workshop on Language Technologies for Historical and Ancient Languages*, European Language Resources Association (ELRA), Paris, pp. 124-129.

Sychev, O. and Penskoy, N.A. (2019), *Method of lemmatizer selections in multiplexing lemmatization*, in «IOP Conference Series Materials Science and Engineering», 483, pp. 1-6.

Tarrant, R. (2016), *Texts, Editors, and Readers: Methods and Problems in Latin Textual Criticism*, Cambridge University Press, Cambridge.

Van Dam, H.-J. (1982), *A laughing Jabberwocky*, in «Wauwelwok: The Magazine of Het Nederlands Lewis Carroll Genootschap», 5, pp. 6-13.

Whitaker, W. (1993), *Words v.1.97F* [available online at *http://archives.nd.edu/whitaker/wordsdoc.htm*, accessed on 21.05.2020].

Wu, W. and Nicolai, G. (2020), *JHUBC's submission to LT4HALA EvaLatin 2020*, in Sprugnoli, R. and Passarotti, M. (2020, eds.), *Proceedings of 1st Workshop on Language Technologies for Historical and Ancient Languages*, European Language Resources Association (ELRA), Paris, pp. 114-118.

Zeman, D., Hajič, J., Popel, M., Potthast, M., Straka, M., Ginter, F., Nivre, J. and Petrov, S. (2018), *CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies*, in Hajič, J. and Zeman, D. (2018, eds.), *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Association for Computational Linguistics, Stroudsburg, PA, pp. 1-21.

Patrick J. Burns
Department of Classics
University of Texas at Austin
WAG 123
Austin, TX 78712 (United States)
patrick.burns@austin.utexas.edu

SSL

Interlinking through lemmas. The lexical collection of the *LiLa* Knowledge Base of linguistic resources for Latin

Marco Passarotti, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo, Rachele Sprugnoli

Abstract

This paper presents the structure of the *LiLa* Knowledge Base, i.e. a collection of multifarious linguistic resources for Latin described with the same vocabulary of knowledge description and interlinked according to the principles of the so-called Linked Data paradigm. Following its highly lexically based nature, the core of the *LiLa* Knowledge Base consists of a large collection of Latin lemmas, serving as the backbone to achieve interoperability between the resources, by linking all those entries in lexical resources and tokens in corpora that point to the same lemma. After detailing the architecture supporting *LiLa*, the paper particularly focusses on how we approach the challenges raised by harmonizing different strategies of lemmatization that can be found in linguistic resources for Latin. As an example of the process to connect a linguistic resource to *LiLa*, the inclusion in the Knowledge Base of a dependency treebank is described and evaluated.

Keywords: linguistic resources, linguistic linked open data, lemmatization, interoperability, Latin.

1. *Introduction*

Linguistic resources are machine-readable collections of language data and descriptions typically divided into two categories depending on the kind of content they include: (i) textual resources, such as written and spoken corpora, featuring either partial or full texts of various typologies, which may differ in genre, author or time period, and (ii) lexical resources, for instance lexica, dictionaries and terminological databases, providing information on lexical items for one or more languages, including definitions, translations and morphological properties. In most cases, linguistic resources do not only feature data, namely texts and lists of lexical items, but also metadata, which enhance the resource with a medley of annotations ranging from descriptive information (e.g. structural division into books, chapters, etc.) to linguistic

Received: January 2020
Accepted: April 2020

traits, such as lemmatization, Part-of-Speech (PoS) tags and syntactic function.

Over the past two decades the research area dedicated to building, improving and evaluating linguistic resources has seen substantial growth and, today, covers a wide span of languages and language varieties. This progress speaks to the need of larger (meta)data sets to support empirically-based studies and to the fact that most (stochastic) systems, tools or algorithms for Natural Language Processing (NLP) currently rely on the linguistic and meta-linguistic evidence stored in corpora or lexica. The strict relation holding between NLP tools and linguistic resources is twofold. On the one hand, NLP tools exploit the empirical data provided by resources to build trained models, whose accuracy rates heavily depend on the size (and quality) of the training data. On the other, the development of new resources, as well as the extension of existing ones, is supported by NLP tools, which automatically enrich (textual or lexical) data with linguistic metadata.

Despite the increase in the quantity and coverage of linguistic resources, most of these are locked in data silos, which prevent users from honing both their individual and joint potential in interoperable ways. While resources tend to focus on providing annotation at one or more levels of linguistic analysis – be those lexical, morphological, syntactic, semantic or pragmatic – linking them to one another helps to draw the overall picture and to maximize their individual contribution. Indeed, linguistic data and metadata today are scattered in distributed resources, thus failing to provide a comprehensive overview of the annotations available in these separate collections. One of the main challenges at the present time is interlinking the motley amount of linguistic data and metadata stored in the resources developed over the past five decades of Computational Linguistics and empirical language studies (Chiarcos *et al.*, 2012: 1). Overcoming this challenge is no simple task because: (a) linguistic resources are often designed for particular tasks (e.g. PoS tagging and syntactic analysis); (b) linguistic resources and NLP tools may use different conceptual models (e.g. different PoS tagsets); (c) linguistic data might be represented using different formalisms (e.g. annotation schemas), which are often incompatible between systems (van Erp, 2012: 58).

We owe this predicament to the fact that, throughout the years, more attention has been given to making linguistic resources grow in size, complexity and diversity, rather than making them interact. Tentative solutions

to the problem of resource isolation, such as the CLARIN[1], DARIAH[2] and META-SHARE[3] linguistic infrastructures and databases, are but upshots of the last decade. What these initiatives provide, however, is a single query access point to multiple meta-collections of resources and tools, rather than connections between them. Instead, making linguistic resources interoperable requires that all types of annotation applied to a particular word/text be integrated into a common representation for indiscriminate access to any linguistic information provided by a resource or tool (Chiarcos, 2012: 162).

A current approach to interlinking linguistic resources takes up Linked Data principles, so that «it is possible to follow links between existing resources to find other, related data and exploit network effects» (Chiarcos *et al.*, 2013: iii). According to the Linked Data paradigm, data in the Semantic Web (Berners-Lee *et al.*, 2001) are interlinked through connections that can be semantically queried, so as to make the structure of web data better serve the needs of users. In the area of linguistic resources, the Linguistic Linked Open Data cloud (*LLOD*)[4] is a collaborative effort pursued by several members of the Open Linguistics Working Group[5], with the general goal of developing a Linked Open Data (sub-)cloud of linguistic resources (McCrae *et al.*, 2016). Indeed, the application of Linked Data to linguistic data ultimately connects Linguistics to other domains that have adopted the paradigm, including Geography (Goodwin *et al.*, 2008), Biomedicine (Ashburner *et al.*, 2000) and Government[6].

What this fervent area of research still lacks, however, is a fine-grained level of interaction between linguistic resources capable of stretching beyond descriptive metadata over to individual word occurrences in a text or entries in a lexicon.

One subfield that has enjoyed particular prosperity over the past decade is that devoted to ancient languages. Owing to their key role in accessing and understanding the so-called Classical tradition, Latin and Ancient Greek are among the main beneficiaries.

Although Latin was among the first languages to be automatically processed with computers thanks to the pioneering work on the texts of

[1] Cf. *https://www.clarin.eu/*.
[2] Cf. *https://www.dariah.eu/*.
[3] Cf. *http://www.meta-share.org/*.
[4] Cf. *http://linguistic-lod.org/llod-cloud*.
[5] Cf. *https://linguistics.okfn.org/index.html*.
[6] Cf. *https://data.gov.uk/*.

Thomas Aquinas by the Italian Jesuit Roberto Busa in the 1940s, throughout its sixty-year history Computational Linguistics has been mainly focusing on living languages, whose commercial and social impact is larger than that of their dead counterparts. In 2006, however, the launch of two independent (but related) projects aimed at building the first syntactically annotated corpora (called 'treebanks') for Latin brought about a research renaissance of linguistic resources and NLP tools for ancient languages (Bamman *et al.*, 2008). This came as no surprise given the vast amount of texts written in Latin spread all over Europe and covering a time span of almost two millennia. These texts bear testament to a common yet diverse past and have contributed to shaping European cultural heritage. Making full use of the most advanced techniques for preserving, investigating and sharing this legacy is at the same time a challenge and an obligation for the research community.

Thanks to international efforts, several textual and lexical resources, as well as NLP tools, are currently available for Latin. Despite the launch of a number of projects for automatic extraction of structured knowledge from ancient sources in the last decade (see Section 2), much like other languages, linguistic resources and tools for Latin often live in isolation, a condition which prevents them from benefiting a large research community of historians, philologists, archaeologists and literary scholars.

To this end, the *LiLa*: Linking Latin project (2018-2023)[7] was awarded funding from the European Research Council (*ERC*) to build a Knowledge Base of linguistic resources for Latin based on the Linked Data paradigm, i.e. a collection of multifarious, interlinked data sets described with the same vocabulary of knowledge description (by using common data categories and ontologies). The project's ultimate goal is to make full use of the linguistic resources and NLP tools for Latin developed thus far, in order to bridge the gap between raw language data, NLP and knowledge description (Declerck *et al.*, 2012: 111).

This paper presents the structure of the lexical basis of *LiLa*, which serves as the backbone of the Knowledge Base to achieve interoperability between textual and lexical resources for Latin. Following a summary of the linguistic resources currently available for Latin (Section 2), we detail the architecture supporting *LiLa*, with special focus on how we approach the challenges raised by harmonizing different strategies of lemmatization

[7] Cf. *https://lila-erc.eu*.

(Section 3). The inclusion in *LiLa* of a dependency treebank is described and evaluated in Section 4. Finally, Section 5 discusses a number of open questions to be addressed by the project in the near future.

2. *Linguistic resources for Latin*

A wealth of linguistic resources is digitally available for Latin today as a result of decades' worth of work spent turning paper-based textual and lexical data into machine-readable formats. This section seeks to provide a brief overview of these efforts to delineate the quantity and diversity of the linguistic data currently at our disposal.

With regard to textual resources, among the most prominent collections of digital texts are the Perseus Digital Library[8], the corpus of Latin texts developed by the Laboratoire d'Analyse Statistique des Langues Anciennes (*L.A.S.L.A.*)[9], the Bibliotheca Teubneriana Latina by De Gruyter[10], the collection of Classical Latin texts prepared by the Packard Humanities Institute (*PHI*)[11], the Loeb Classical Library[12], and a set of collections published by Brepols, such as the Library of Latin Texts[13], the Archive of Celtic Latin Literature[14] and the Aristoteles Latinus Database[15]. More recently, the Digital Latin Library project[16] set out to publish and curate critical editions of Latin texts of all types, genres and eras. A similar objective is pursued by the Open Greek and Latin project[17], whose ultimate goal is to represent every source text produced in Classical Greek or Latin in Antiquity (through c. 600 AD) with a view to covering also the Post-classical era until modern times. The project places the total number of Ancient Greek and Latin words surviving from Antiquity at 150 million, and the number of Post-classical Latin words available in some 10,000 books in the Internet Archive at 200 million.

[8] Cf. *http://www.perseus.tufts.edu/hopper/*.
[9] Cf. *http://web.philo.ulg.ac.be/lasla/*.
[10] Cf. *https://www.degruyter.com/view/db/btl*.
[11] Cf. *http://latin.packhum.org/*.
[12] Cf. *https://www.loebclassics.com/*.
[13] Cf. *https://about.brepolis.net/library-of-latin-texts/*.
[14] Cf. *http://www.brepols.net/Pages/BrowseBySeries.aspx?TreeSeries=ACLL-O*.
[15] Cf. *http://www.brepols.net/Pages/BrowseBySeries.aspx?TreeSeries=ALD*.
[16] Cf. *https://digitallatin.org/*.
[17] Cf. *http://www.opengreekandlatin.org/*.

By virtue of infrastructural efforts conducted over the past decade, this large amount of textual resources is now accessible via aggregating initiatives, including Corpus Corporum[18], a meta-collection containing more than 150 million words in texts written in Ancient Greek or Latin provided by more than twenty different corpora and collections; Trismegistos[19], a portal of papyrological and epigraphical resources formerly covering Egypt and the Nile valley (800 BC-800 AD) and now expanding to the Ancient World in general; and the eAqua project[20], conceived to support the search of co-occurrences and citations in a number of collections of Ancient Greek and Latin texts, including Perseus and *PHI*.

Beside these catchall (meta)collections comprising large number of texts, genres and authors diachronically spread from Antiquity to Neo-Latin, some corpora provide more specific data. The Patrologia Latina database[21], for instance, features more than 100 million words from the writings of the Church Fathers; the Musisque Deoque digital archive[22] contains poetic works by some 200 authors; late-antique Latin texts are made available by the *digilibLT* Digital Library, which currently boasts 265 works written before the 6th century AD[23]; the Corpus Grammaticorum Latinorum[24] gathers the *Grammatici Latini*, that is, Latin grammar manuals written between the 2nd and 7th centuries AD and edited by Heinrich Keil in Leipzig from 1855 to 1880. As for Medieval Latin, the Index Thomisticus by father Roberto Busa SJ (Busa, 1974-1980)[25] collects the opera omnia of Thomas Aquinas, for a total of over 11 million words, the *ALIM* corpus[26] provides texts of the Italian Latinity of the Middle Ages, and the Computational Historical Semantics project[27] is a large database of Medieval Latin texts from various sources.

Among other distinctive digital corpora for Latin, noteworthy examples are the School of Salamanca[28], a digital text corpus of 116 works of Sal-

[18] Cf. *http://www.mlat.uzh.ch/MLS/*.
[19] Cf. *https://www.trismegistos.org/index.php*.
[20] Cf. *http://www.eaqua.net/*.
[21] Cf. *http://pld.chadwyck.co.uk/*.
[22] Cf. *http://mizar.unive.it/mqdq/public/*.
[23] Cf. *http://digiliblt.lett.unipmn.it/index.php*.
[24] Cf. *https://bibliotheque.univ-paris-diderot.fr/bases-de-donnees/cgl-corpus-grammaticorum-latinorum*.
[25] Cf. *http://www.corpusthomisticum.org/*.
[26] Cf. *http://www.alim.dfll.univr.it/*.
[27] Cf. *https://www.comphistsem.org/home.html*.
[28] Cf. *https://www.salamanca.school/en/works.html*.

mantine jurists and theologians found in selected printed books published between the 16th-17th centuries; the *CroALa* corpus brings together some 450 writings by 181 Croatian Latin authors, for a total of over 5 million words produced between the 10th and 20th centuries[29], the Domus sermonum compilatorium archive[30] provides the texts of the sermons of the Franciscan preacher Osvladus de Lasko; the Roman Inscriptions of Britain[31] hosts multiple corpora, including the Vindolanda tablets; Epistolae[32] is a collection of medieval Latin letters written between the 4th and 13th centuries to and from women; DanteSearch[33] provides both the vernacular and the Latin writings of Dante Alighieri, the Latin portion of the corpus counting approximately 46,000 words; finally, *CLaSSES*[34] is a collection of more than 1,200 non-literary Latin texts, such as epigraphs and letters, from different eras (between the 4th century BC and the 6th century AD) and sources (Rome, Central Italy, Britain, Egypt and the Eastern Mediterranean Sea).

A subset of the Latin texts carries linguistic annotation. The most common layer of linguistic annotation available in Latin corpora is lemmatization, which in some cases is also enriched with PoS and morphological tagging. For instance, while the data provided by *CLaSSES* and Roman Inscriptions from Britain are lemmatized, the large collection of texts assembled by *L.A.S.L.A.*, the Index Thomisticus, DanteSearch, as well as roughly one million tokens of the Computational Historical Semantics corpus are all fully lemmatized and morphologically tagged.

Syntactic annotation, on the other hand, is still limited to a small set of texts. Four treebanks are currently available for Latin. These are: (i) the Index Thomisticus Treebank (*IT-TB*) (Passarotti, 2019), based on the works of Thomas Aquinas; (ii) the Latin Dependency Treebank (*LDT*) (Bamman and Crane, 2006) of texts belonging to the Classical era, now part of the Ancient Greek and Latin Dependency Treebank 2.0 under development at the University of Leipzig (Celano, 2019); (iii) the *PROIEL* corpus (Pragmatic Resources in Old Indo-European Languages), which features the syntactic annotation of the oldest extant versions of the New Testament in Indo-European languages and Latin texts from

[29] Cf. *http://www.ffzg.unizg.hr/klafil/croala/*.
[30] Cf. *http://sermones.elte.hu/szovegkiadasok/latinul/laskaiosvat/index.php?file=os_index*.
[31] Cf. *https://romaninscriptionsofbritain.org/*.
[32] Cf. *https://epistolae.ctl.columbia.edu/*.
[33] Cf. *http://www.perunaenciclopediadantescadigitale.eu:8080/dantesearch/*.
[34] Cf. *http://classes-latin-linguistics.fileli.unipi.it/*.

both the Classical and Late eras (Haug and Jøhndal, 2008); and (iv) the Late Latin Charter Treebank (*LLCT*), a syntactically annotated corpus of original 8th-9th century charters from Central Italy (Korkiakangas and Passarotti, 2011). While the *LDT*, the *IT-TB* and the *LLCT* have shared the same syntactic annotation schema since their inception (Bamman *et al.*, 2007), resembling that of the so-called analytical layer of annotation of the Prague Dependency Treebank for Czech (Hajič *et al.*, 1999), the *PROIEL* treebank follows a slightly different style (Haug, 2010). At present, with the exception of the *LLCT*, all Latin treebanks are also available in the Universal Dependencies collection (*UD*) (Nivre *et al.*, 2016)[35]. In terms of size, the *IT-TB* currently counts some 350,000 annotated words, *LDT* counts 55,000, the Latin section of the *PROIEL* corpus 200,000 and *LLCT* counts 250,000 annotated words.

With regard to lexical resources, among the many dictionaries and lexica available in digital format today are the Lewis and Short dictionary accessible through Perseus, the Thesaurus Linguae Latinae of the Bayerische Akademie der Wissenschaften in Munich[36], and Johann Ramminger's Neulateinische Wortliste[37]. Brepols provides an extensive list of Latin word forms, known as Thesaurus Formarum Totius Latinitatis[38], with number of occurrences for each in the Library of Latin Texts, and the comprehensive Database of Latin Dictionaries[39], which itself consists of a large number of different types of lexical resources. Another noteworthy initiative is Logeion[40], a cross-dictionary search tool, providing simultaneous lookup of entries in the many lemmatized works from the Perseus Classical collection by way of the PhiloLogic system[41]. Within the Computational Historical Semantics project there is the Frankfurt Latin Lexicon, a lexical resource built upon assorted source lexicons and taggers and used for NLP tasks, such as morphological tagging, lemmatization, and PoS tagging[42].

The availability of Latin treebanks has made it possible to induce subcategorization lexica from the *IT-TB* (*IT-VaLex*) (McGillivray and Passarotti, 2009) and the *LDT* (*VaLex*) (McGillivray, 2013). Latin Vallex is a valency

[35] Cf. *https://universaldependencies.org/.*
[36] Cf. *https://www.degruyter.com/view/db/tll.*
[37] Cf. *http://www.neulatein.de/.*
[38] Cf. *http://www.brepols.net/Pages/BrowseBySeries.aspx?TreeSeries=TF.*
[39] Cf. *https://about.brepolis.net/database-of-latin-dictionaries/.*
[40] Cf. *https://logeion.uchicago.edu/.*
[41] Cf. *http://philologic.uchicago.edu/.*
[42] Cf. *https://www.comphistsem.org/lexicon0.html.*

lexicon built in conjunction with the semantic and pragmatic annotation of the *IT-TB* and the *LDT* (Passarotti *et al.*, 2016). Presently, Latin Vallex includes around 1,350 lexical entries. The LatinWordNet (*LWN*) (Minozzi, 2010) was built in the context of the MultiWordNet project (Pianta *et al.*, 2002), whose aim was to build a number of semantic networks for specific languages aligned with the synsets of the Princeton WordNet (*PWN*) (Fellbaum, 2012)[43]. The language-specific synsets were created by translating *PWN* data with the help of bilingual dictionaries. The *LWN* counts 8,973 synsets and 9,124 lemmas, and is currently undergoing substantial revision with a view to refining and extending its contents (Franzini *et al.*, 2019). The Word Formation Latin (*WFL*) lexicon (Litta and Passarotti, 2019) provides information about derivational morphology by connecting lemmas via word formation rules[44].

LiLa seeks to maximize the use of these (and many other) resources for Latin by making them interoperable, thus allowing users to run complex queries across linked and distributed resources, like, for instance, searching the four Latin treebanks for occurrences of verbs featuring a specific (a) dependency relation, e.g. subject (source: treebanks), (b) prefix (source: *WFL*), (c) valency frame (source: Latin Vallex), and (d) belonging to a particular WordNet synset (source: *LWN*).

3. *The* LiLa *Knowledge Base*

In this section we describe the architecture of the *LiLa* Knowledge Base, built to structure the information of the Latin linguistic resources in a centralized hub of interaction.

In order to achieve interoperability between distributed resources, *LiLa* makes use of a set of Semantic Web and Linked Data standards and practices. These include ontologies to describe linguistic annotation (*OLiA*: Chiarcos and Sukhareva, 2015), corpus annotation (NLP Interchange Format (*NIF*): Hellmann *et al.*, 2013; *CoNLL-RDF*: Chiarcos and Fäth, 2017) and lexical resources (Lemon: Buitelaar *et al.*, 2011; Ontolex: McCrae *et al.*, 2017).

[43] Synsets are unordered sets of cognitive synonyms, i.e. words that denote the same concept and are interchangeable in many contexts. In WordNets, synsets are interlinked by means of conceptual-semantic and lexical relations.

[44] Cf. *http://wfl.marginalia.it/*.

Following Bird and Liberman (2001), the Resource Description Framework (*RDF*) (Lassila and Swick, 1998) is used to encode graph-based data structures to represent linguistic annotations in terms of triples: (i) a predicate-property (a relation; in graph terms: a labeled edge) that connects (ii) a subject (a resource; in graph terms: a labeled node) with (iii) its object (another resource/node, or a literal, e.g. a string). The *SPARQL* language is used to query the data recorded in the form of *RDF* triples (Prud'Hommeaux and Seaborne, 2008).

3.1. *Linking through lemmatization*

Lemmatization is a layer of annotation and organization of linguistic data common to different kinds of resources. Dictionaries tend to index lexical entries using lemmas. Thesauri organize the lexicon by collecting all related entries, and use lemmas to index them; so, for instance, the nominal synset n#07202206 of the *PWN*, glossed as "a female human offspring", is lexicalized in *LWN* by the lemmas: *filia* "daughter", *nata* "daughter" and *puella* "girl". Lemmas are also used to facilitate lexical search in corpora. This is particularly helpful for languages, like Latin, with rich inflectional morphology; a regular Latin verb, for instance, can have up to 130 forms (if we exclude the nominal inflection of the participles or gerundives), with varying endings and, at times, different stems.

Given the presence and role played by lemmatization in various linguistic resources, and the good accuracy rates achieved by the best performing lemmatizers for Latin (up to 95.30%, as per Eger *et al.*, 2015)[45], *LiLa* uses the lemma as the most productive interface between lexical resources, annotated corpora and NLP tools. Consequently, the *LiLa* Knowledge Base is highly lexically based, grounding on a simple, but effective assumption that strikes a good balance between feasibility and granularity: textual resources

[45] Such high rates of automatic lemmatization of Latin should be taken with a grain of salt. Indeed, performances of stochastic NLP tools heavily depend on the training set on which their models are built, and so decrease when they are applied to out-of-domain texts. This problem is particularly challenging for Latin owing to its wide diachrony (spanning two millennia), genre diversity (ranging from literary to philosophical, historical and documentary texts) and diatopy (Europe and beyond). For the state of the art in automatic lemmatization and PoS tagging for Latin, see the results of the first edition of *EvaLatin*, a campaign devoted to the evaluation of NLP tools for Latin (Sprugnoli *et al.*, 2020). The first edition of *EvaLatin* focused on two shared tasks (i.e. lemmatization and PoS tagging), each featuring three sub-tasks (i.e. Classical, Cross-Genre, Cross-Time). These sub-tasks were specifically designed to measure the impact of genre variation and diachrony on NLP tool performances.

are made of (occurrences of) words, lexical resources describe properties of words, and NLP tools process words.

Figure 1 presents the main components of the *LiLa* Knowledge Base, showing the key interlinking role played by the Lemma node. A 'Lemma' is an (inflected) 'Form' chosen as the citation/canonical form of a lexical item. Lemmas occur in 'Lexical Resources' as citation/canonical forms of lexical entries. Forms, too, can occur in lexical resources, like in a lexicon containing all of the forms of a language (for instance, Tombeur, 1998). Both Lemmas and Forms can have 'Morphological Features', such as PoS, gender, mood and tense. The occurrences of Forms in real texts are 'Tokens', which are provided by 'Textual Resources'. Finally, on NLP tools performances can process either Textual Resources (e.g. a tokenizer), Forms, regardless of their contextual use (e.g. a morphological analyzer), or Tokens (e.g. a PoS tagger).

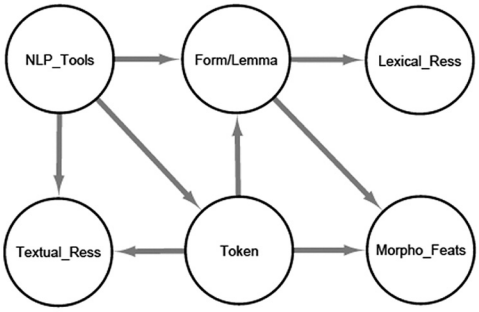

Figure 1. *The main components of* LiLa.

The core of the *LiLa* Knowledge Base consists of a large collection of Latin lemmas: interoperability is achieved by linking all those entries in lexical resources and tokens in corpora that point to the same lemma. While the process of selecting the canonical forms to be used as lemmas tends to follow a standardized series of language-dependent conventions (e.g. for Latin, the nominative singular form for nouns, or the first person singular of the active indicative present tense for verbs), building and structuring a repository of canonical forms that may serve as a hub in *LiLa* is complicated by the fact that different corpora, lexica and tools adopt different strategies to solve the conceptual and linguistic challenges posed by lemmatization, namely (a) the form of the lemma and (b) lemmatization criteria.

Citation forms for the same lexical item chosen to represent the lemma differ in (a) graphical representation (*voluptas* vs *uoluptas* "satisfaction"), (b)

spelling (*sulfur* vs *sulphur* "brimstone"), (c) ending and possibly inflectional type (*diameter* vs *diametros* vs *diametrus* "diameter"), or (d) in the paradigmatic slot representing the lemma (*sequor* "to follow", first person singular of the passive/deponent present indicative vs *sequo*, first person singular of the active present indicative). Furthermore, homographic lemmas, like *occīdo* (*ob+caedo* "to strike down") and *occĭdo* (*ob+cado* "to fall down"), can either be left ambiguous by using the same character string *occido* for the forms of both lemmas, or told apart. For instance, in the Index Thomisticus corpus *occīdo* and *occĭdo* are recorded as *occido^caedo* and *occido^cado*, respectively, while in the *LDT* (and in the Perseus Digital Library in general) as *occido1* and *occido2*.

As for lemmatization criteria, differences are such that, on occasion, a word form can be reduced to multiple lemmas. This is the case of participles, which can be considered either as part of the verbal inflectional paradigm or as independent lemmas deserving of a separate entry in lexical resources. Accordingly, participles can either be lemmatized under the main verb or under a dedicated participial lemma, which in turn may be used either systematically or only when the participle has grown into an autonomous lexical item (e.g. *doctus* "learned", morphologically the perfect participle of *doceo* "to teach"). The same holds true for deadjectival adverbs (e.g. *aequaliter* "evenly" from *aequalis* "equal"), which are either lemmatized as forms of their base adjective, as happens in the *IT-TB*, or treated as independent lemmas, like in the *PROIEL* treebank. Another issue is raised by polythematic words for which missing forms are taken from other stems, as is the case of *melior* used as the comparative of *bonus* (see English "good" and "better"). These are sometimes subsumed under the (positive degree of the) adjective or given a self-standing lemma.

3.2. *The* LiLa *ontology of Latin canonical forms*

Cases like the disambiguation of the ambiguous forms *occīdo* and *occĭdo* attest to the variety of lemmatization solutions different resources may adopt. In this respect, it is important to note that the approach of *LiLa* is not to harmonize resources by choosing one lemmatization standard over another or by imposing prescriptive guidelines to which all lemmatized resources must be converted. Rather, *LiLa* aims to provide a descriptive set of concepts and properties capable of integrating *all* solutions adopted by different Latin resources.

To this end, *LiLa* implements a formal ontology, expressed in the Web Ontology Language (*OWL*; McGuinness and Van Harmelen, 2004), that defines the classes, properties and instances involved in the task of lemmatization, as well as the possible interactions between lemmas, lemmatized corpora and lexica. Since the ultimate goal of the project is to establish a network of linguistic resources fully interoperable within the *LLOD* cloud, this ontology reuses as many existing standards as possible. In this way, we ensure that the data amassed by *LiLa* are immediately compatible with other Linked (Open) Data resources.

The *LiLa* ontology starts by defining the class of the Lemma, the pivotal concept in our domain. In our definition, lemmatization is the task of indexing all inflected forms under one that is conventionally identified as canonical. As such, the Lemma is safely subsumed under the general class of Form as defined in the Ontolex ontology, a *de facto* standard in the Linked Data publication of lexical resources. Relying on the concepts of Ontolex, we define the Lemma as a Form that is linked to a Lexical Entry via the property 'canonical form'. This structural choice allows us to potentially connect all other lexical resources compiled using the Ontolex (or Lemon) formalism to our collection.

Forms are grammatical realizations of words or of any other class of Lexical Entries that have at least one written representation. The Ontolex 'written representation' property can be used to accommodate the different spellings or peculiar inflections of canonical forms: in the case of the examples discussed above, *sulfur* and *sulphur* become two written representations of the same lemma, and so do the loan words that display either the Greek or the Latin endings (like *diametros* and *diametrus*)[46]. We, therefore, use this property whenever the variation in the realization of a lemma affects only the orthography of a form (including the word ending), provided that its morphological analysis and the inflectional paradigm are not altered.

What Ontolex also permits is the inclusion of a phonetic representation of a form. As vocalic quantity is often used to disambiguate between homographic words (again, *occīdo* and *occĭdo*), we add a special sub-property for prosodic representation, which carries all the relevant transcriptions of a form with long and short vowel diacritics. The variation, however, may involve changes in PoS, inflectional paradigm or other morphological features.

[46] But note that if the variation also entails a different type of inflection (such as *diameter* on the one hand and *diametrus/diametros* on the other), we represent the lemmas as two different forms linked to one another via the property 'lemma variant' (see below).

Some Latin words belong to more than one PoS, as is, for example, the case of prepositions that can be used as adverbs. Since Lexical Entries in Ontolex cannot have more than one PoS[47], the same restriction applies also to canonical forms. Accordingly, *LiLa* will provide two lemmas with written representation *ante* "before", one for the preposition and one for the adverb.

Participles and inflectional variation are harder to model and require an extension of the Ontolex ontology. Some words present two or more alternative inflectional paradigms, which entail different lemmas. Verbs with both a deponent and an active inflection, for example, are often found in Latin lexica. Although one of the paradigms might be more frequent and more 'regular' than another from a traditional lexicography or grammar standpoint[48], we cannot exclude that corpora in which the 'irregular' instances are met lemmatize these under the less typical canonical form. As a consequence, *LiLa* records all possible canonical forms as lemmas; so, in our collection, the verbs *sequor*[49] and *sequo*[50], for example, exist as independent lemmas. Since these forms can both be used to lemmatize instances of the same words, we link them to one another with the symmetric property 'lemma variant', thus making it possible to retrieve from the textual resources connected to *LiLa* all the tokens that belong to the same lexical item, regardless of the lemmatization criteria followed in individual corpora.

Participles, again, behave differently. As previously mentioned, participles like *docti* "learned" can be reduced to a form of either *doceo* "to learn" or *doctus* "learned". In these cases, that is, whenever a form can be interpreted as part of the (regular) inflectional paradigm or as a Lemma in itself, we associate that form to a special sub-class of Lemma called Hypolemma. Hyper- and hypolemmas are linked to one another via the symmetric property 'has hypolemma'/'is hypolemma'[51].

A Lemma is also defined by a series of morphological features. All lemmas are assigned a PoS (which, as we have already seen, must be exclusive for each form), and can be analyzed by those traits that are typical of nominal (gender, number, case), adjectival (gender, number, case, degree) and verbal

[47] See the definition of Lexical Entry at *https://www.w3.org/2016/05/ontolex/#lexical-entries*.

[48] In the case of verbs *sequor*/*sequo*, the active form *sequo* is mentioned by grammarians only: see Gell. 18.9.8 and Prisc. *Ars Gramm.* 9.28.

[49] Cf. *https://lila-erc.eu/lodview/data/id/lemma/124461*.

[50] Cf. *https://lila-erc.eu/data/id/lemma/124462*.

[51] Note that, with respect to its hyperlemma, a hypolemma entails a change in the PoS: *faciliter* "easily" is an adverb, while *facilis* "easy" is an adjective; *doctus* (as an autonomous lemma) is an adjective, while *doceo* is a verb.

(tense, mood, person, number, voice) inflection; additionally, lemmas have an inflectional type (i.e. the conjugations and declensions of traditional grammars). *LiLa*'s ontology formalizes these linguistic properties together with the relevant restrictions, so that, for instance, tense cannot be predicated of nouns. The PoS tags adopted in *LiLa* are based on the universal tagset of Universal Dependencies (Petrov *et al.*, 2011). However, in order to ensure compatibility with other tagsets used for Latin, *LiLa*'s categories for linguistic annotation are aligned with the *OLiA* ontology. So, for instance, *LiLa*'s class 'Adjective' is a sub-class of *OLiA*'s 'Adjective', which also subsumes all other tags used to annotate the same grammatical category.

Lemmas can also be analyzed in terms of their derivational morphology. This level integrates the information recorded in the *WFL* lexicon into the *LiLa* collection. Since an Ontolex extension for derivational morphology is currently under development, this module is still not available for immediate deploying. Ontolex allows lexical resources to describe derivational morphemes as regular lexical entries, provided with written representations. However, for our ontology, we opted for a minimal extension only. In *LiLa*, morphemes belong to their own class, and are grouped into Affixes (distinguishing between prefixes and suffixes) and Bases. We define the Base as the lexical morpheme of a word that is neither a prefix nor a suffix. Words that are derived, even in several steps, from the same root (for instance, *adduco* "to lead to", *adductio* "bringing in", *duco* "to lead", *produco* "to lead forth" and *productivus* "productive") are therefore linked to the same base.

This conceptual architecture was first put to the test with a comprehensive list of Latin canonical forms based on the one provided by the Latin morphological analyzer *Lemlat* (Passarotti *et al.*, 2017), which was used to populate the *LiLa* collection[52]. *Lemlat*'s database reconciles three reference dictionaries for Classical Latin (*GGG*: Georges and Georges, 1913-1918; Glare, 1982; Gradenwitz, 1904)[53], the entire *Onomasticon* from Forcellini's (1940) *Lexicon Totius Latinitatis* (Budassi and Passarotti, 2016) and the Medieval Latin *Glossarium Mediae et Infimae Latinitatis* by du Cange *et al.* (1883-1887), for a total of over 150,000 lemmas (Cecchini *et al.*, 2018b).

[52] Cf. *https://github.com/CIRCSE/LEMLAT3*.

[53] The choice of lexicographic sources for Classical Latin in *Lemlat* is based on the remarks by Lomanto (1980).

The linguistic properties of these lemmas are expressed as *RDF* triples using the *LiLa* ontology formalism and are stored in a triplestore publicly accessible via a *SPARQL* endpoint[54]. *Lemlat*'s lemmas have undergone a twofold process of revision: firstly, we removed overlapping or duplicate lemmas between the Classical and Medieval forms; secondly, we generated hypolemmas for all the canonical forms of present, future and perfect participles, as well as for deadjectival adverbs, and connected them to their main hyperlemmas via the symmetric property 'has hypolemma'/'is hypolemma'.

The *LiLa* collection currently includes 130,925 lemmas, 92,947 hypolemmas, 292,657 written representations of (hypo)lemmas, 59,945 'has/is hypolemma' properties, and 6,120 links between lemma variants[55].

3.3. *Examples from the lexical collection of* LiLa

In this section, we report on examples taken from the Knowledge Base to show the way in which a lemma and its connected information are stored in the *LiLa* lexical collection. More specifically, we detail how lemma variants, morphological features, hypolemmas, information on derivational morphology and prosodic representations are treated.

We first consider the lemma *claudeo*/*claudo* "to limp". In the *Oxford Latin Dictionary* (Glare, 1982), the entry for this lemma includes both the second conjugation (*claudeo*) and third conjugation verbs (*claudo*), the latter also featuring the graphical variant *cludo* (Lucil. 250). The lemma is recorded as deriving from the first class adjective *clausus* "closed, inaccessible".

In the *Ausführliches Lateinisch-Deutsches Handwörterbuch* (Georges and Georges, 1913-1918), alongside the citation forms *claudeo* and *claudo* we also find their respective and semantically identical deponent counterparts *claudeor* and *claudor*.

In the du Cange Medieval Latin *Glossarium*, lexical entries are provided neither for *claudeo*/*-eor* nor *claudo*/*-or*.

As previously mentioned, the *Lemlat* lexical basis integrates the *GGG* dictionaries. In *Lemlat*, the information about *claudeo*/*claudo* provided by these three reference dictionaries is merged into one single entry; here, a

[54] Cf. *https://lila-erc.eu/sparql/*. A network-based access point to the collection is available at *https://lila-erc.eu/lodlive/* and a user-friendly query interface is accessible at *https://lila-erc.eu/query/*.

[55] Numbers subject to change as the process of elimination of duplicate lemmas is still ongoing.

common ID is assigned to all lexical bases used to build the citation forms of the lexical entry. In the example case, *Lemlat* contains five different citation forms for the same lexical entry, all bearing the same ID: *claudeo*, *claudeor*, *claudo*, *claudor*, and *cludo*. In *LiLa*, these citation forms are represented by four lemmas distinguished by inflectional category. *Claudeo* and *claudo*, as well as their corresponding deponent forms *claudeor* and *claudor*, are citation forms for different lemmas, as they follow two different inflectional categories (active and deponent second conjugation, respectively)[56]. *Cludo*, on the other hand, is merged with *claudo*, as these share the same inflection. Just like *sequor* and *sequo*, *LiLa* connects these four lemmas via the 'lemma variant' property, while *cludo* and *claudo* are represented as different written representations, i.e. graphical variants of the same lemma (Figure 2)[57].

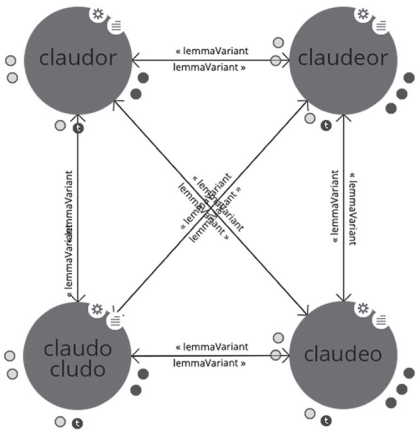

Figure 2. *Four citation forms of the same lexical entry in* LiLa.

In doing so, and as previously mentioned, *LiLa* harmonizes different lemmatization strategies and annotation styles, thus granting interoperability. In the example of *claudeo/claudo*, all the tokens of this lexical item occur-

[56] The homographic lemma of the third conjugation *claudo* "to close" is an independent node in *LiLa*, separate from *claudeo/claudo* and, thus, given a different unique identifier in the Knowledge Base.

[57] In all *LiLa* Figures henceforth (taken from the Lodlive interface), the small 'satellite' nodes circling the larger ones represent links to other nodes in the Knowledge Base, e.g. the PoS of the lemma.

ring in the lemmatized corpora and lexica available in *LiLa* can be joined together by using a set of five connected citations, regardless of whether the citation form used in a specific textual resource is *claudeo*, *claudeor*, *claudor*, or *claudo*/*cludo*.

The criterion used to distinguish between the different citation forms and different written representations of the same lexical item is purely morphological and, specifically, inflectional. If two citation forms for the same item belong to different inflectional categories, they are considered (and thus represented in *LiLa*) as two separate lemmas connected via the 'lemma variant' property. If not, they are stored in the lexical collection of *LiLa* as two written representations of the same lemma. Indeed, each Lemma node in *LiLa* is connected to a number of morphological features, among which is the inflectional category, as indicated by the 'has inflection type' property. Figure 3 shows the different categories to which the possible citation forms for *claudeo/claudo* are connected.

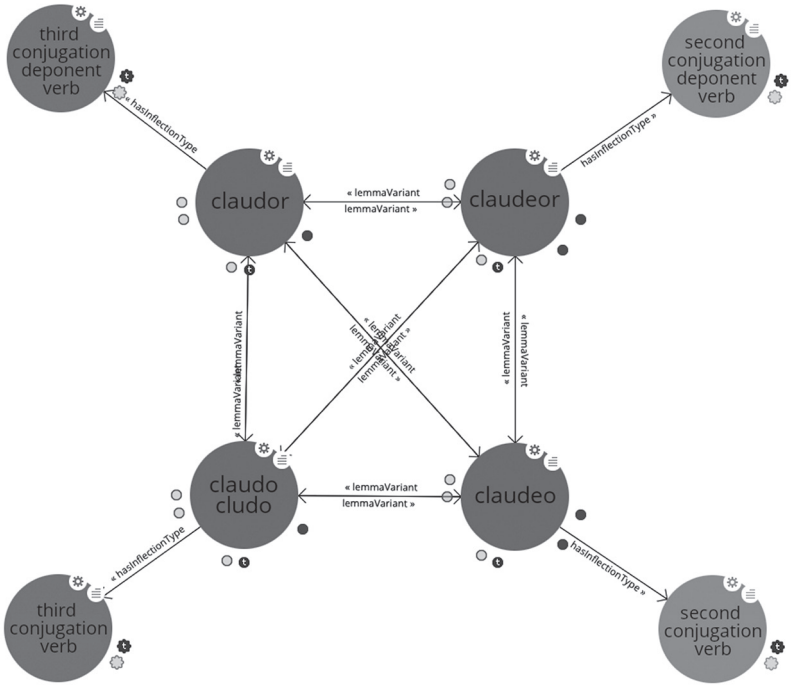

Figure 3. *Inflectional categories in* LiLa.

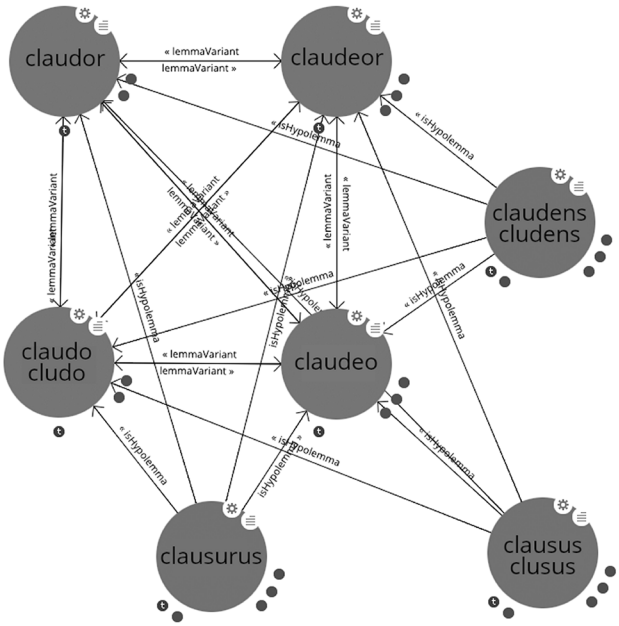

Figure 4. *Hypolemmas of verbs in* LiLa.

As we have already seen, Lemma nodes in *LiLa* can be connected to those for hypolemmas. In the case of lemmas for verbs, these are all connected to their hypolemmas for present, future and perfect participles. As Figure 4 shows, the node for *claudeo* (lemma) is connected to those for its participial citation forms *clausurus/clusurus*, *claudens/cludens* and *clausus/clusus* (hypolemmas) via the relation 'is hypolemma', making it possible to join different lemmatization strategies for participles. The same holds true for the other three lemmas connected via 'lemma variant'. In this way, whether in a lemmatized corpus a form like *claudentem* is assigned lemma *claudeo* (or *claudo*, *cludo*, *claudor*, *claudeor*) or *claudens*, in *LiLa* the form is always connected to the same lemma, as *claudens* is the written representation of the hypolemmas of all four lemmas for *claudeo/claudo*. Once again, *LiLa* does not perform any analysis but merely reflects the disambiguation provided by the connected resources. This means that, be it assigned to *claudo* or *claudens*, the form *claudentem* in *LiLa* is connected to both *claudeo/claudo* and *claudo* "to close". If the source corpus (or lexicon) includes morphological annotation,

the connection of the form to the correct lemma can be partly disambiguated on the basis of inflection, seeing as *claudeo* and *claudo* belong to two different categories[58]. Instead, if the resource to be included in *LiLa* does not provide morphological annotation but lemmatization and PoS tagging only, any form associated with the lemma *claudo* or *claudens* would be connected to both *claudeo/claudo* and *claudo*.

Beside inflectional morphological features, *LiLa* lemmas also carry information on derivational morphology. Two types of information about word formation are provided. Firstly, all lemmas belonging to a derivational family, i.e. a set of (derived) lemmas sharing the same lexical base, are connected to a node common to all family members (Base)[59]. Secondly, lemmas formed with one or more derivational affixes are connected to the nodes for such affixes (prefixes or suffixes). The information on derivational morphology is taken from the *WFL* resource by flattening the hierarchical relations of derivation recorded therein. Indeed, while *WFL* represents derivational families in terms of rooted trees, where one lemma is hierarchically derived from another (or from others, in the case of compounds), *LiLa* does not include such hierarchical relations between lemmas, but represents derivational morphology via flat connections between lemmas and their base(s) and affix(es) (Litta *et al*., 2019). Figure 5 shows the derivational family tree of *claudeo* in *WFL*.

In the derivational tree of Figure 5, each node represents a lemma belonging to the same derivational family. Nodes are connected by hierarchical relations labelled with the respective word formation rule. For instance, the lemma *claudeo/-eor* is the result of an adjective-to-verb conversion rule (A-To-V) applied to the adjective *claudus* "limping". The verb *claudico* "to limp", in turn, is derived from *claudeo/-eor* as a deverbal verb with the suffix *-ic*.

Like *LiLa*, *WFL* too makes use of the *Lemlat* lexical basis and so inherits the tool's lemma merges (e.g. *claudeo/-eor*). In *LiLa*, however, *claudeo* and *claudeor* are separate lemmas connected via the property 'lemma variant'. Furthermore, *LiLa* uses 'lemma variant' also to connect the third conjugation lemmas *claudo/cludo* and *claudor*; these are missing from *WFL* despite being recorded in *Lemlat* as variant forms of *claudeo/-eor*. Figure 6 shows how the derivational family of *claudeo* is represented in *LiLa*.

[58] This disambiguation is only partial. In order to disambiguate between *claudo* "to limp" and *claudo* "to close" (both third conjugation verbs) the resource must provide additional information other than morphology, e.g. a reference to the semantics of the lexical item.

[59] Compounds are connected to more than one Base node.

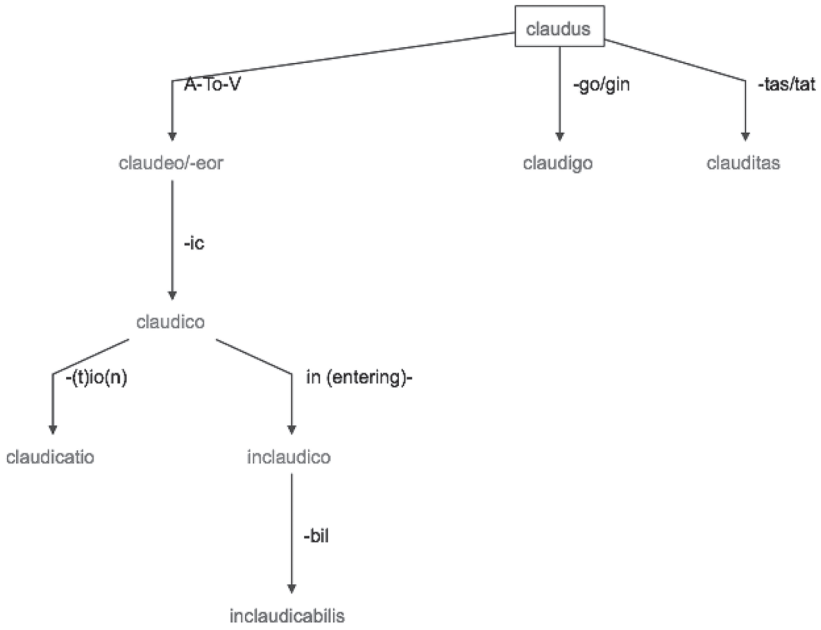

Figure 5. *The derivational family tree of* claudeo *in* WFL.

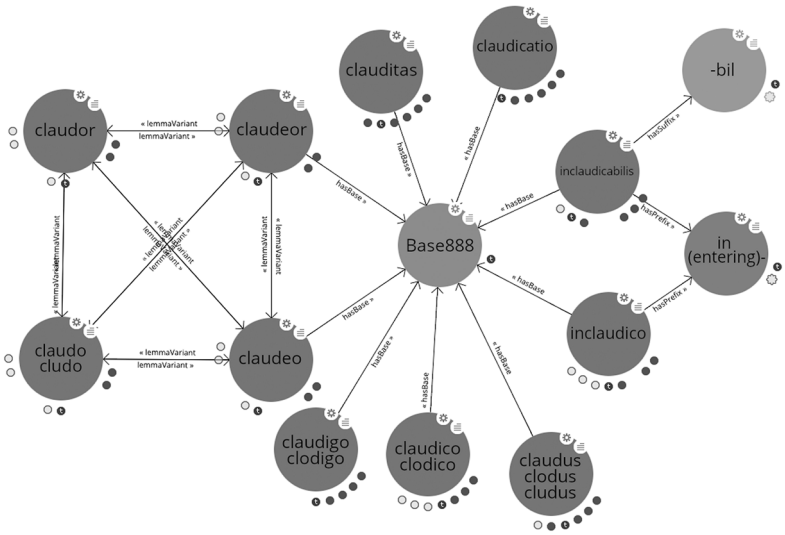

Figure 6. *The derivational family of* claudeo *in* LiLa.

In Figure 6, each lemma of the derivational family of *claudeo* is connected to a common Base node via the relation 'has Base'. As a connector between lemmas of a family, the Base node is unspecific and is instead given a numeric label (in this case, 888)[60]. Those lemmas that include one or more affixes are connected to the nodes for such affixes via the 'has prefix' and 'has suffix' properties, respectively. In Figure 6, this is the case of *inclaudico* "to limp / to be lame" and *inclaudicabilis* "not limping": while both lemmas are connected to the prefix node *in (entering)-* via the relation 'has prefix', *inclaudicabilis* alone is connected to the suffix node *-bil* via the 'has suffix' relation. Since the lemma variants *claudo*/*cludo* and *claudor* do not occur in *WFL* but in *LiLa* only, they are not explicitly connected to the Base node 888. These relations, however, are automatically induced in the ontology of *LiLa* in that all lemmas connected via 'lemma variant' share, possibly via inheritance, the same base and affixes (where present).

As mentioned in Section 3.2, cases like *occīdo* vs *occĭdo* are handled by attaching a 'prosodic representation' with vowel length to the lemma. Figure 7 shows the representation in *LiLa* of the verb *occīdo*.

The lemma node for the verb *occīdo* (with Type 'Lemma'), is connected to (a) its participial hypolemmas (*occisurus*, *occisus* and *occidens*), (b) its PoS ('Verb'), (c) the prefix *ob-*, (d) the inflection type 'third conjugation verb' and (e) Base 37, which is shared with, for instance, the verb *peroccido*, "to kill thoroughly". Moreover, the node *occīdo* is connected to the written representation *occido* and to the prosodic representation *occīdo*.

[60] Base nodes lack any kind of explicitly recorded linguistic information, as doing so would require a clear definition of the linguistic status of Base nodes stretching beyond that of connectors between lemmas belonging to the same derivational family. Indeed, such definition would open up a number of issues. One possible solution could be to assign each Base node a written representation consisting of a string describing the lexical 'element' (a root? a stem?) underlying each lemma in the derivational family (e.g. *dic-* for *dico* "to say", or *dictio* "a saying"). This procedure is complicated by the fact that different bases can be used in the same family, as is the case of, for example, *fer-*, *tul-* and *lat-*, which can all be found as bases in the family to which the verb *fero* "to bring" belongs. However, the current treatment of Base nodes does not prevent from integrating etymological information in the *LiLa* Knowledge Base (Mambrini and Passarotti, 2020).

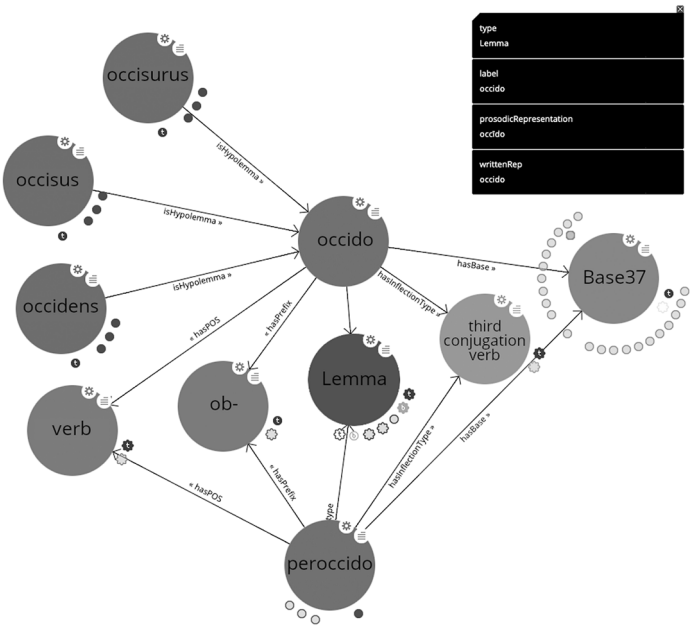

Figure 7. *Prosodic Representation in* LiLa.

4. *Including linguistic resources into* LiLa

Compiling the collection of lemmas described in previous sections is not the ultimate objective of *LiLa*, but a necessary step towards achieving interoperability between the linguistic resources included in the Knowledge Base.

In metaphysical terms, the collection of lemmas in *LiLa* represents a set of *noumena* (and it is, in itself, a *noumenon*), and a resource is a provider of *phenomena* (and it is, in itself, a *phenomenon*). The definition of these terms in Webster's Online Dictionary reads[61]:

The *noumenon* (plural: *noumena*) classically refers to an object of human inquiry, understanding or cognition. The term is generally used in contrast with, or in relation to, *phenomenon* (plural: *phenomena*), which refers to appearances, or objects

[61] Cf. *http://www.websters-dictionary-online.org*.

of the senses. A *phenomenon* is that which is perceived; A *noumenon* is the actual object that emits the *phenomenon* in question.

In *LiLa*, lemmas exist regardless of their actual realizations in textual and/or lexical resources. The first step of the *LiLa* project was to build this 'lexical *noumenon*'. The second step is to connect the *noumenon* to the *phenomenon*, i.e. to its actual realizations.

So far, the only textual resource to have been connected to *LiLa* is the *IT-TB* in its original annotation schema. This section describes the process of connecting the *IT-TB* to *LiLa* and details how the (meta) data provided by this treebank are linked to the lemma collection of the Knowledge Base.

The *IT-TB* exists in *LiLa* in its version downloadable from the *IT-TB* website (December, 2019)[62]. This version includes a selection of the concordances of the lemma *forma* "form" extracted from three works of Thomas Aquinas and the full text of the first three books of the *Summa contra gentiles*, for a total of 277,547 tokens (239,496 lexical tokens and 38,051 punctuation marks), corresponding to 3,901 different lemmas[63].

To connect the lemmatized lexical tokens of the *IT-TB* to the *LiLa* collection of lemmas, we perform a simple string match between the lemmas in the treebank and the written representations of lemmas in the Knowledge Base. As a result of this strategy, 3,627 out of 3,901 lemmas in the *IT-TB* (corresponding to 233,291 lexical tokens) were linked to at least one lemma in *LiLa*, while 274 (corresponding to 6,205 lexical tokens) found no match. Out of 3,697 lemmas, 778 were linked ambiguously[64] or, in other words, connected to more than one lemma in *LiLa*; in *LiLa*, for example, there exist two lemmas with written representation *venio*, both of which are verbs, one first conjugation ("to genuflect", a rare Medieval word from the du Cange glossary) and the other fourth conjugation ("to come")[65].

[62] Cf. *https://itreebank.marginalia.it*.

[63] Details on the composition of the *IT-TB* can be found in Passarotti (2019).

[64] Unambiguous linking obtained through simple string match may be risky in the case of homographic lemmas missing from the *LiLa* lexical collection, i.e. when a lemma in the incoming resource is a homograph of only one written representation of a lemma in *LiLa*, but belongs to another homographic lemma not present in the collection.

[65] The integration in *LiLa* of lexical resources providing information like, for instance, the date of first attestation of a lemma, its frequency, or its prevalence in a specific genre, will help to reduce ambiguity in the linking process.

To disambiguate cases like *venio*, we use the morphological tagging provided by the *IT-TB*, which assigns to each word form its PoS and inflectional category (declension, conjugation)[66]. For instance, in the sentence (1):

(1) *Nam primo habet formam seminis, postea sanguinis, et sic inde quousque veniat ad ultimum complementum.* (Thom. *Summa contra gentiles* II 89,9)
"At first it possesses the form of semen, afterwards of blood, and so on, until at last it arrives at that wherein it finds its fulfilment." [67]

the word form *veniat* in the *IT-TB* is assigned the PoS 'Verb' and the fourth conjugation, thus making it possible to unambiguously link it to the correct lemma in *LiLa*. This strategy disambiguated 650 lemmas out of the ambiguous 778 previously linked.

This leaves us with 128 ambiguously linked lemmas, because the lemmatization and morphological tagging of the *IT-TB* preclude an automated choice between the candidate lemmas. This is the case of the lemma *campus* (a second declension masculine noun), which links to *campus* "field" and *campus* (*marinus*) for *hippocampus* "sea-horse".

Finally, a number of lemmas were still left unlinked. These were found to fall under one of the following categories:

– the lemma does not exist in the *LiLa* collection, as is the case of the third declension feminine noun *actualitas* "actuality" (as opposed to potentiality). The *IT-TB* counts 223 of these cases, besides which 4 are new hypolemmas (e.g. the adverb *quantum* "as much as" recorded as hypolemma of *quantus* "how much") and 24 are lemmas of the type *occido^caedo*/*occido^cado*, for which disambiguation was performed manually: *IT-TB* tokens connected to *occido^caedo* were linked to the lemma with prosodic representation *occīdo*, while those connected to *occido^cado* were linked to *occĭdo*;

– the lemma of the *IT-TB* is a new written representation of a lemma already present in *LiLa*; this is the case of the written representation *annuncio* for the first conjugation verb *adnuntio* "to announce". Eight cases;

– the lemma of the *IT-TB* is a new lemma variant of a lemma already present in *LiLa*. For example, the singular first declension masculine noun

[66] PoS tagging in the *IT-TB* does not make use of the usual PoS labels, but follows three inflectional classes: nominal inflection (for nouns, adjectives and pronouns, with a separate tag for the nominal forms of the verbal paradigms: gerunds, gerundives, participles and supines), verbal inflection (for verbs) and no inflection (for adpositions, adverbs, conjunctions and interjections). Further details on the tripartite tagging of the *IT-TB* can be found in Cecchini *et al.* (2018a).

[67] English translation from *https://dhspriory.org/thomas/english/ContraGentiles2.htm#89*.

anthropomorphita is a lemma variant of the corresponding pluralia tantum *anthropomorphitae* (a group of heretics who attributed human form to God). Three cases;

– so-called 'pseudo-lemmas', which are used in the *IT-TB* for non Latin words (*non latina vox*), numbers (*num. arab.* and *num. rom.* for Arabic and Roman numbers, respectively) and abbreviations (e.g. *breviata loci notatio*). Eleven cases;

– lemmatization errors in the *IT-TB*. Six cases, e.g. *pbiectum* instead of *obiectum* "object".

After classifying lemmas into these categories, we expanded the *LiLa* collection with the new lemmas, written representations and lemma variants needed to fully connect the *IT-TB* to *LiLa*[68]. This strategy exemplifies *LiLa*'s empirical approach, whereby the lexical basis of the Knowledge Base grows with the number of linguistic resources connected.

The syntax of the *IT-TB* is annotated in dependency trees. Figure 8 shows the *IT-TB* dependency tree of sentence (1).

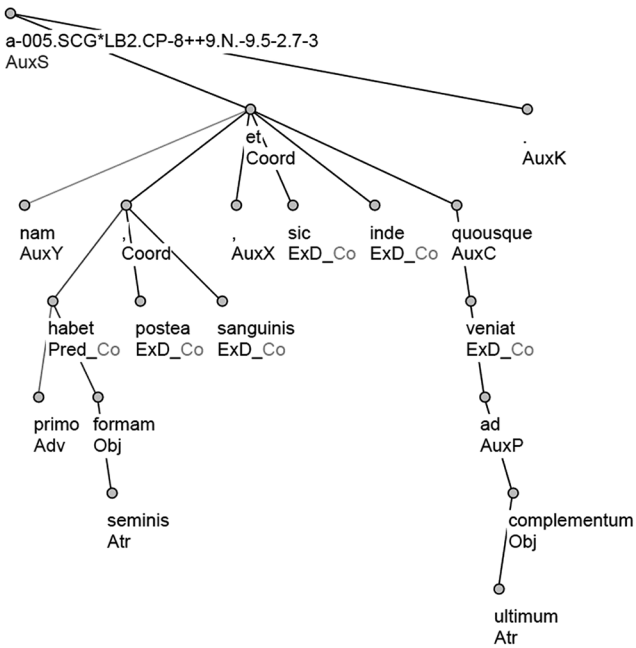

Figure 8. *A dependency tree from the Index Thomisticus Treebank.*

[68] Pseudo-lemmas and lemmatization errors remain unlinked.

The tree in Figure 8 features as many nodes as there are tokens in the sentence, including punctuation. Each token is assigned a syntactic function, known in dependency treebank jargon as 'dependency relation' (DepRel)[69]. Figure 9 shows the graphical representation of the connections holding between the tokens of the clause *quousque veniat ad ultimum complementum* (part of sentence 1) and the lemmas in the *LiLa* collection.

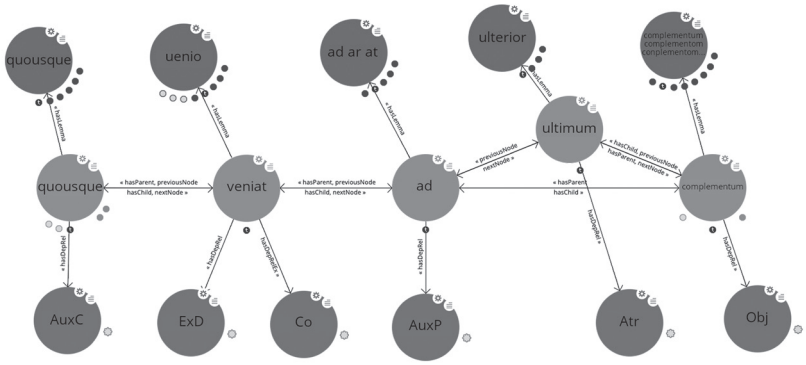

Figure 9. *A clause of the Index Thomisticus Treebank in* LiLa.

In Figure 9, each token of the example clause in the *IT-TB* is connected to exactly one lemma in *LiLa* via the relation 'has lemma', and to its previous/next node in the sentence via the symmetric relation 'previous node'/'next node'[70].

In the *LiLa* Knowledge Base, two pieces of information can be extracted from the trees of a dependency treebank:

(i) tokens are connected to their syntactic function via the property 'has DepRel'. The dependency relations shown in Figure 9 are AuxC (for subordinating conjunctions, here *quousque*), ExD (for nodes missing their head node in the dependency tree, i.e. ellipsis, here *veniat*), AuxP

[69] For a detailed description of the annotation rules and the set of dependency relations used in the *IT-TB*, see Bamman *et al.* (2007).

[70] Each token is also connected to a number of descriptive metadata taken from the original linguistic resource. In the case of the *IT-TB*, each token is linked to descriptive metadata recording its position in the texts of Thomas Aquinas (e.g. work, book, chapter, etc.) and to the sequence of morphological tags originally attached to it in the *IT-TB* (e.g. 3-MB1--6--1 for the third person singular of the present subjunctive of fourth conjugation verbs, e.g. *veniat*). The full morphological tagset of the *IT-TB* is available at *https://itreebank.marginalia.it/doc/Tagset_IT.pdf*.

(for prepositions, here *ad*), Atr (for Attributes, here *ultimum*) and Obj (for direct/indirect objects, i.e. arguments, here *complementum*). In the *IT-TB*, the syntactic functions of nodes in coordinated constructions are indicated by the extension _Co, as evidenced by *veniat* in Figure 8. In *LiLa*, this is represented via the relation 'has DepRelEx', which in Figure 9 connects the token *veniat* to the node *Co*;

(ii) dependencies between head and dependent nodes are represented through the symmetric property 'has parent'/'has child'. In Figure 9, for instance, the relation 'has child' holding between *veniat* and *ad* indicates that *veniat* is the head of *ad* in the dependency tree of this *IT-TB* clause[71].

5. *Conclusion*

In this paper, we have presented the overall architecture of the *LiLa* Knowledge Base of linguistic resources for Latin. Interweaving the large amount of linguistic (meta)data developed thus far in an interoperable whole is key to promoting the use of resources and tools. Today, this is made possible thanks to Linked Data technologies.

The first objective of the *LiLa* project was to compile a large collection of Latin lemmas in Linked Data form. This collection, described here in Section 3, represents the backbone of *LiLa*, given the central role played by the lemma in making resources interact. The collection was derived from a number of reference dictionaries and glossaries covering different chronological eras. However, as demonstrated by the inclusion of the first linguistic resource in the Knowledge Base (the Index Thomisticus Treebank; Section 4), a complete lexical coverage is far from being achieved (if not impossible), seeing as future resources are expected to introduce new lexical items and/or new citation forms of lemmas already recorded in *LiLa*. The greater the number of resources connected in *LiLa*, the larger its lemma collection will become.

The important role of the lemma in *LiLa* implies that only lemmatized resources can fully exploit the (lexical) connections in the Knowledge Base. Nowadays, this is a restrictive condition as, despite growing numbers, many Latin corpora do not carry this layer of linguistic annotation. One core chal-

[71] When the 'has parent'/'has child' property overlaps with the 'previous node'/'next node' one, these are merged into one edge in the visualization, as exemplified by *veniat* and *ad* in Figure 9: *veniat* both precedes *ad* in the word order of the clause and it is its parent node in the dependency tree.

lenge for *LiLa* will be to collect and evaluate the tools and trained models available for automatic lemmatization and, next, to build a new set to allow data providers to process their resource(s) for ready inclusion in the Knowledge Base. Indeed, even if lemmatized, texts might nevertheless cause trouble in cases such as ambiguous homographic lemmas (e.g. *occīdo* vs *occĭdo*). *LiLa*, after all, reflects the degree of annotation granularity provided by the resources attached to the Knowledge Base.

Another important issue that *LiLa* must address is how to deal with resources in closed and/or proprietary formats. While most Computational Linguistics resources and tools are freely available, popular collections of scholarly editions of Latin and Ancient Greek texts, such as the Bibliotheca Teubneriana Latina by De Gruyter and all Brepols corpora, are locked behind paywalls. In line with the 'as open as possible, as closed as necessary' approach, proprietary resources will be connected in the Knowledge Base but access to them will be subject to charges. In doing so, we hope to influence policy change and to establish *LiLa* as a leading publication venue of Latin's linguistic legacy.

Acknowledgments

We are thankful to Daniela Corbetta and Andrea Peverelli for their invaluable support in building and extending the *LiLa* collection of lemmas. The *LiLa*: Linking Latin project has received funding from the European Research Council (*ERC*) under the European Union's Horizon 2020 research and innovation programme - Grant Agreement No 769994.

References

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. (2000), *Gene ontology: tool for the unification of biology*, in «Nature genetics», 25, 1, p. 25.

Bamman, D. and Crane, G. (2006), *The design and use of a Latin dependency treebank*, in Hajič, J. and Nivre, J. (2006, eds.), *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories*, Institute of Formal and Applied Linguistics, Prague, pp. 67-78.

Bamman, D., Passarotti, M., Busa, R. and Crane, G. (2008), *The annotation guidelines of the Latin Dependency Treebank and Index Thomisticus Treebank. The treatment of some specific syntactic constructions in Latin*, in Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S. and Tapias, D. (2008, eds.), *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, European Language Resources Association (ELRA), Paris, pp. 71-76.

Bamman, D., Passarotti, M., Crane, G. and Raynaud, S. (2007), *Guidelines for the syntactic annotation of Latin treebanks*, Tufts University Digital Library, Medford / Somerville.

Berners-Lee, T., Hendler, J. and Lassila, O. (2001), *The Semantic Web*, in «Scientific American», 284, 5, pp. 28-37.

Bird, S. and Liberman, M. (2001), *A formal framework for linguistic annotation*, in «Speech communication», 33, 1-2, pp. 23-60.

Budassi, M. and Passarotti, M. (2016), Nomen Omen. *Enhancing the Latin morphological analyser* Lemlat *with an onomasticon*, in Reiter, N., Alex, B. and Zervanou, K.A. (2016, eds.), *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2016)*, Association for Computational Linguistics, Berlin, pp. 90-94.

Buitelaar, P., Cimiano, P., McCrae, J., Montiel-Ponsoda, E. and Declerck, T. (2011), *Ontology lexicalisation: The lemon perspective*, in Slodzian, M., Valette, M., Aussenac-Gilles, N., Condamines, A., Hernandez, N. and Rothenburger, B. (2011, eds.), *Proceedings of the Workshops. 9th International Conference on Terminology and Artificial Intelligence*, INALCO, Paris, pp. 33-36.

Busa, R. (1974-1980), *Index Thomisticus: sancti Thomae Aquinatis operum omnium indices et concordan-tiae, in quibus verborum omnium et singulorum formae et lemmata cum suis frequentiis et contextibus variis modis referuntur quaeque consociata plurium opera atque electronicoIBM automato usus digessit Robertus Busa SJ.*, Frommann / Holzboog, Stuttgart / Bad Cannstatt.

DU CANGE, C., BÉNÉDICTINS DE SAINT-MAUR, CARPENTIER, P., HENSCHEL, L. and FAVRE, L. (1883-1887), *Glossarium Mediae et Infimae Latinitatis*, L. Favre, Niort.

CECCHINI, F.M., PASSAROTTI, M., MARONGIU, P. and ZEMAN, D. (2018a), *Challenges in converting the Index Thomisticus Treebank into Universal Dependencies*, in DE MARNEFFE, M.C., LYNN, T. and SCHUSTER, S. (2018, eds.), *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, The Association for Computational Linguistics, Bruxelles, pp. 27-36.

CECCHINI, F.M., PASSAROTTI, M., TESTORI, M., RUFFOLO, P., DRAETTA, L., FIEROMONTE, M., LIANO, A., MARINI, C. and PIANTANIDA, G. (2018b), *Enhancing the Latin morphological analyser* LEMLAT *with a Medieval Latin glossary*, in CABRIO, E., MAZZEI, A. and TAMBURINI, F. (2018, eds.), *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, Accademia university press, Torino, pp. 87-92.

CELANO, G.G.A. (2019), *The Dependency Treebanks for Ancient Greek and Latin*, in BERTI, M. (2019, ed.), *Digital Classical Philology. Ancient Greek and Latin in the Digital Revolution*, De Gruyter, Berlin / Boston, pp. 279-298.

CHIARCOS, C. (2012), *Interoperability of corpora and annotations*, in CHIARCOS, C., HELLMANN, S. and NORDHOFF, S. (2012, eds.), *Linked Data in Linguistics*, Springer, Berlin, pp. 161-179.

CHIARCOS, C., HELLMANN, S. and NORDHOFF, S. (2012), *Introduction and overview*, in CHIARCOS, C., HELLMANN, S. and NORDHOFF, S. (2012, eds.), *Linked Data in Linguistics*, Springer, Berlin, pp. 1-12.

CHIARCOS, C., CIMIANO, P., DECLERCK, T. and MCCRAE, J.P. (2013), *Linguistic linked open data (*llod*). Introduction and overview*, in CHIARCOS, C., CIMIANO, P., DECLERCK, T. and MCCRAE, J.P. (2013, eds.), *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and Linking Lexicons, Terminologies and Other Language Data*, Association for Computational Linguistics, Pisa, pp. i-xi.

CHIARCOS, C. and SUKHAREVA, M. (2015), OLiA - *Ontologies of Linguistic Annotation*, in «Semantic Web Journal», 6, 4, pp. 379-386.

CHIARCOS, C. and FÄTH, C. (2017), CoNLL-RDF*: Linked corpora done in an* NLP*-friendly way*, in GRACIA, J., BOND, F., MCCRAE, J., BUITELAAR, P., CHIARCOS, C. and HELLMANN, S. (2017, eds.), *Language, Data, and Knowledge*, Springer, Berlin, pp. 74-88.

Declerck, T., Lendvai, P., Mörth, K., Budin, G. and Váradi, T. (2012), *Towards linked language data for digital humanities*, in Chiarcos, C., Hellmann, S. and Nordhoff, S. (2012, eds.), *Linked Data in Linguistics*, Springer, Berlin, pp. 109-116.

Eger, S., vor der Brück, T. and Mehler, A. (2015), *Lexicon-assisted tagging and lemmatization in Latin: A comparison of six taggers and two lemmatization methods*, in Zervanou, K., van Erp, M. and Alex, B. (2015, eds.), *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, Association for Computational Linguistics, Beijing, pp. 105-113.

Fellbaum, C. (2012), *Wordnet*, in Chapelle, C. (2012, ed.), *The Encyclopedia of Applied Linguistics*, Wiley Online Library [doi:*10.1002/9781405198431.wbeal1285*, accessed on 28.11.2019].

Forcellini, E. (1940), *Lexicon Totius Latinitatis / ad Aeg. Forcellini lucubratum, dein a Jos. Furlanetto emendatum et auctum; nunc demum Fr. Corradini et Jos. Perin curantibus emendatius et auctius meloremque in formam redactum adjecto altera quasi parte Onomastico totius latinitatis opera et studio ejusdem Jos. Perin*, Typis Seminarii, Padova.

Franzini, G., Peverelli, A., Ruffolo, P., Passarotti, M., Sanna, H., Signoroni, E., Ventura, V. and Zampedri, F. (2019), *Nunc Est Aestimandum: Towards an Evaluation of the Latin WordNet*, in Bernardi, R., Navigli, R. and Semeraro, G. (2019, eds.), *Proceedings of the Sixth Italian Conference on Computational Linguistics* (AI*IA Series, 2481), CEUR Workshop Proceedings, Bari, pp. 1-8.

Georges, K.E. and Georges, H. (1913-1918), *Ausführliches Lateinisch-Deutsches Handwörterbuch*, Hahn, Hannover.

Glare, P.G.W. (1982), *Oxford Latin Dictionary*, Oxford University Press, Oxford.

Goodwin, J., Dolbear, C. and Hart, G. (2008), *Geographical linked data: The administrative geography of great britain on the semantic web*, in «Transactions in GIS», 12, pp. 19-30.

Gradenwitz, O. (1904), *Laterculi Vocum Latinarum*, Hirzel, Leipzig.

Hajič, J., Panevová, J., Buráňová, E., Urešová, Z. and Bémová, A. (1999), *Annotations at analytical level. Instructions for annotators*, UK MFF ÚFAL, Prague [available online at *https://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/a-layer/pdf/a-man-en.pdf*, accessed on 28.11.2019].

Haug, D.T.T. and Jøhndal, M. (2008), *Creating a parallel treebank of the old Indo-European Bible translations*, in Sporleder, C. and Ribarov, K. (2008, eds.), *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, European Language Resources Association (ELRA), Paris, pp. 27-34.

Haug, D. (2010), Proiel *guidelines for annotation* [available online at *http://folk.uio.no/daghaug/syntactic_guidelines.pdf*, accessed on 28.11.2019].

Hellmann, S., Lehmann, J., Auer, S. and Brümmer, M. (2013), *Integrating NLP using Linked Data*, in Alani, H., Lalana, K., Fokoue, A., Groth, P., Biemann, C., Xavier Parreira, J., Aroyo, L., Noy, N., Welty, C. and Janowicz, K. (2013, eds.), *The Semantic Web – ISWC 2013. 12th International Semantic Web Conference, Sydney, Australia, October 21-25, 2013*, Springer, Berlin / Heidelberg, pp. 98-113.

Korkiakangas, T. and Passarotti, M. (2011), *Challenges in annotating medieval Latin charters*, in «Journal for Language Technology and Computational Linguistics», 26, 2, pp. 103-114.

Lassila, O. and Swick, R.R. (1998), *Resource description framework (rdf) model and syntax specification* [available online at *https://www.w3.org/TR/1999/REC-rdf-syntax-19990222/*, accessed on 28.11.2019].

Litta, E. and Passarotti, M. (2019), *(When) inflection needs derivation: a word formation lexicon for Latin*, in Holmes, N., Ottink, M., Schrickx, J. and Selig, M. (2019, eds.), *Lemmata Linguistica Latina*. Vol. 1: *Words and Sounds*, De Gruyter, Berlin / Boston, pp. 224-239.

Litta, E., Passarotti, M. and Mambrini, F. (2019), *The treatment of word formation in the* LiLa *Knowledge Base of linguistic resources for Latin*, in Žabokrtský, Z., Ševčíková, M., Litta, E. and Passarotti, M. (2019, eds.), *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2019). 19-20 September 2019, Prague*, Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Prague, pp. 35-43.

Lomanto, V. (1980), *Lessici latini e lessicografia automatica*, in «Memorie dell'Accademia delle Scienze di Torino. Classe di Scienze Morali, Storiche e Filologiche», 5, 4, pp. 113-270.

Mambrini, F. and Passarotti, M. (2020), *Representing etymology in the* LiLa *knowledge base of linguistic resources for Latin*, in Kernerman, I. and Krek, S. (2020, eds.), *Proceedings of the Globalex Workshop on Linked Lexicography (@LREC 2020)*, European Language Resources Association (ELRA), Paris.

McCrae, J.P., Chiarcos, C., Bond, F., Cimiano, P., Declerck, T., De Melo, G., Gracia, J., Hellmann, S., Klimek, B., Moran, S., Osenova, P., Pareja-Lora, A. and Pool, J. (2016), *The open linguistics working group: Developing the Linguistic Linked Open Data cloud*, in Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J. and Piperidis, S. (2016, eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, European Language Resources Association (ELRA), Paris, pp. 2435-2441.

McCrae, J.P., Bosque-Gil, J., Gracia, J., Buitelaar, P. and Cimiano, P. (2017), *The Ontolex-Lemon model: development and applications*, in Kosem, I., Tiberius, C., Jakubíček, M., Kallas, J., Krek, S. and Baisa, V. (2017, eds.), *Proceedings of eLex 2017 conference*, Lexical Computing, Brno, pp. 19-21.

McGillivray, B. and Passarotti, M. (2009), *The Development of the Index Thomisticus Treebank Valency Lexicon*, in Borin, L. and Lendvai, P. (2009, eds.), *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH – SHELT &R 2009)*, Association for Computational Linguistics, Athens, pp. 43-50.

McGillivray, B. (2013), *Methods in Latin Computational Linguistics*, Brill, Leiden.

McGuinness, D.L. and Van Harmelen, F. (2004), OWL *web ontology language overview*, in Web Ontology Working Group (2004, ed.), *W3C recommendation*, 10.10 [available online at *https://www.w3.org/TR/2004/REC-owl-features-20040210/*, accessed on 28.11.2019].

Minozzi, S. (2010), *The Latin WordNet project*, in Anreiter, P. and Kienpointner, M. (2010, eds.), *Latin Linguistics Today. Akten des 15. Internationalen Kolloquiums zur Lateinischen Linguistik, Innsbrucker Beiträge zur Sprachwissenschaft*, Institut für Sprachen und Literaturen der Universität Innsbruck Bereich Sprachwissenschaft, Innsbruck, pp. 707-716.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R. and Zeman, D. (2016), *Universal Dependencies v1: A multilingual treebank collection*, in Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J. and Piperidis, S. (2016, eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, European Language Resources Association (ELRA), Paris, pp. 1659-1666.

Passarotti, M. (2019), *The project of the Index Thomisticus Treebank*, in Berti, M. (2019, ed.), *Digital Classical Philology. Ancient Greek and Latin in the Digital Revolution*, De Gruyter, Berlin / Boston, pp. 299-319.

Passarotti, M., González Saavedra, B. and Onambele, C. (2016), *Latin Vallex. A treebank-based semantic valency lexicon for Latin*, in Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J. and Piperidis, S. (2016, eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, European Language Resources Association (ELRA), Paris, pp. 2599-2606.

Passarotti, M., Budassi, M., Litta, E. and Ruffolo, P. (2017), *The* Lemlat *3.0 package for morphological analysis of Latin*, in Bouma, G. and Adesam, Y. (2017, eds.), *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language. 22nd May 2017 Gothenburg*, Linköping University Electronic Press, Linköping, pp. 24-31.

Petrov, S., Das, D. and McDonald, R. (2011), *A Universal Part-of-Speech Tagset*, in «ArXiv Preprint» [available online at *https://arxiv.org/abs/1104.2086*, accessed on 28.11.2019].

Pianta, E., Bentivogli, L. and Girardi, C. (2002), *MultiWordNet: Developing an aligned multilingual database*, in Hamdan, H. and Boubiche, D.E. (2002, eds.), *Proceedings of the First International Conference on Global WordNet*, The Association for Computational Linguistics, Liverpool, pp. 55-63.

Prud'Hommeaux, E. and Seaborne, A. (2008), *Sparql query language for rdf. W3c* [available online at *https://www.w3.org/TR/rdf-sparql-query/*, accessed on 28.11.2019].

Sprugnoli, R., Passarotti, M., Cecchini, F.M. and Pellegrini, M. (2020), *Overview of the EvaLatin 2020 evaluation campaign*, in Sprugnoli, R. and Passarotti, M. (2020, eds.), *Proceedings of the LT4HALA 2020 Workshop - 1st Workshop on Language Technologies for Historical and Ancient Languages, satellite event to the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, European Language Resources Association (ELRA), Paris, pp. 105-110.

Tombeur, P. (1998, ed.), *Thesaurus formarum totius latinitatis a Plauto usque ad saeculum Xxum*, Turnhout, Brepols.

van Erp, M. (2012), *Reusing linguistic resources: Tasks and goals for a linked data approach*, in Chiarcos, C., Hellmann, S. and Nordhoff, S. (2012, eds.), *Linked Data in Linguistics*, Springer, Berlin, pp. 57-64.

Marco Passarotti
Facoltà di Scienze Linguistiche
e Letterature Straniere
Università Cattolica del Sacro Cuore
Largo Gemelli 1
20123 Milano (Italy)
marco.passarotti@unicatt.it

Francesco Mambrini
Facoltà di Scienze Linguistiche
e Letterature Straniere
Università Cattolica del Sacro Cuore
Largo Gemelli 1
20123 Milano (Italy)
francesco.mambrini@unicatt.it

Greta Franzini
Facoltà di Scienze Linguistiche
e Letterature Straniere
Università Cattolica del Sacro Cuore
Largo Gemelli 1
20123 Milano (Italy)
greta.franzini@unicatt.it

Flavio Massimiliano Cecchini
Facoltà di Scienze Linguistiche
e Letterature Straniere
Università Cattolica del Sacro Cuore
Largo Gemelli 1
20123 Milano (Italy)
flavio.cecchini@unicatt.it

Eleonora Litta
Facoltà di Scienze Linguistiche
e Letterature Straniere
Università Cattolica del Sacro Cuore
Largo Gemelli 1
20123 Milano (Italy)
eleonoramaria.litta@unicatt.it

Giovanni Moretti
Facoltà di Scienze Linguistiche
e Letterature Straniere
Università Cattolica del Sacro Cuore
Largo Gemelli 1
20123 Milano (Italy)
giovanni.moretti@unicatt.it

Paolo Ruffolo
Facoltà di Scienze Linguistiche
e Letterature Straniere
Università Cattolica del Sacro Cuore
Largo Gemelli 1
20123 Milano (Italy)
paolo.ruffolo@posteo.eu

Rachele Sprugnoli
Facoltà di Scienze Linguistiche
e Letterature Straniere
Università Cattolica del Sacro Cuore
Largo Gemelli 1
20123 Milano (Italy)
rachele.sprugnoli@unicatt.it

NORME PER GLI AUTORI

Le proposte editoriali (articoli, discussioni e recensioni), redatte in italiano, inglese o altra lingua europea di ampia diffusione, vanno inviate tramite il sistema *Open Journal System* (OJS) collegandosi al sito *http://www.studiesaggilinguistici.it* (ove sono indicate le procedure da seguire), utilizzando due formati: un file pdf anonimo e un file word completo di tutti i dati dell'Autore (indirizzo istituzionale e/o privato, numero telefonico ed e-mail).

Nella redazione della proposta editoriale, gli Autori sono invitati ad attenersi scrupolosamente alle norme redazionali della rivista, disponibili sul sito.

Le proposte di articoli e discussioni dovranno essere corredate da un breve riassunto anonimo in lingua inglese, della lunghezza di circa 15 righe o 1.000 battute (spazi inclusi) e da 3 o 4 parole-chiave che individuino dominio e tema dell'articolo.

I contributi saranno sottoposti alla lettura critica di due *referees* anonimi, e quindi all'approvazione del Comitato Editoriale.

Il contributo accettato per la pubblicazione e redatto in forma definitiva andrà inviato tramite OJS nei tempi indicati dal sistema, sia in formato word che pdf, includendo i font speciali dei caratteri utilizzati.

Edizioni ETS
Palazzo Roncioni - Lungarno Mediceo, 16, I-56127 Pisa
info@edizioniets.com - www.edizioniets.com
Finito di stampare nel mese di luglio 2020