Studi e Saggi Linguistici

I contributi pervenuti sono sottoposti alla valutazione di due revisori anonimi.

All submissions are double-blind peer reviewed by two referees.

Studi e Saggi Linguistici è indicizzato in / *Studi e Saggi Linguistici* is indexed in

ERIH PLUS (European Reference Index for the Humanities and Social Sciences)
Emerging Sources Citation Index - Thomson Reuters
L'Année philologique
Linguistic Bibliography
MLA (Modern Language Association Database)
Scopus

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, Germany.

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