The Annotation of Liber Abbaci, a Domain-Specific Latin Resource

Francesco Grotto[1], **Rachele Sprugnoli**[2], **Margherita Fantoli**[3],
Maria Simi[4], **Flavio Massimiliano Cecchini**[2], **Marco Passarotti**[2]
1. Scuola Normale Superiore, Italy
2. Università Cattolica del Sacro Cuore, Italy
3. KU Leuven, Belgium
4. Università degli Studi di Pisa, Italy
francesco.grotto1@sns.it,
{rachele.sprugnoli,flavio.cecchini,marco.passarotti}@unicatt.it
margherita.fantoli@kuleuven.be, maria.simi@unipi.it

Abstract

The *Liber Abbaci* (13th century) is a milestone in the history of mathematics and accounting. Due to the late stage of Latin, its features and its very specialized content, it also represents a unique resource for scholars working on Latin corpora. In this paper we present the annotation and linking work carried out in the frame of the project *Fibonacci 1202-2021*. A gold-standard lemmatization and part-of-speech tagging allow us to elaborate some first observations on the linguistic and historical features of the text, and to link the text to the Lila Knowledge Base, that has as its goal to make distributed linguistic resources for Latin interoperable by following the principles of the Linked Data paradigm. Starting from this specific case, we discuss the importance of annotating and linking scientific and technical texts, in order to (a) compare and search them together with other (non-technical) Latin texts (b) train, apply and evaluate NLP resources on a non-standard variety of Latin. The paper also describes the fruitful interaction and coordination between NLP experts and traditional Latin scholars on a project requiring a large range of expertise.

1 Introduction

Latin texts have a wide diachronic and diatopic extension that corresponds to a similarly large diversity of the textual genres they represent. Besides literary ones, a huge amount of Latin texts of several different genres can be found spread all over Europe and beyond. An important textual genre is represented by scientific treaties, which in many cases are interesting not only for their contents, but also because of the technical terminology they feature.

This is precisely the case for the *Liber Abbaci* 'the book of the abacus' by Leonardo of Pisa (also known as Fibonacci). Written in the very first years of the 1200s, it is a book on arithmetic promoting a style of calculation based on Arabic numerals without aid of an abacus. *Fibonacci 1202-2021* is a project financed by the Tuscany Region and involving the University of Pisa and the Galilei Museum in Florence, following the publication of a critical edition of the *Liber Abbaci* by Enrico Giusti (Fibonacci, 2020). The goal of the project is to produce an enhanced digital edition of this work by leveraging advanced publishing tools and investigating the use of computational linguistics techniques in order to uncover the wealth of linguistic, scientific and historical information contained in the book.

Besides its scientific interest, the *Liber Abbaci* features a very peculiar lexicon, not often represented in the currently available (linguistically annotated) corpora for Latin. In order to fill this gap, in the context of the project *Fibonacci 1202-2021* we have started performing the linguistic annotation of the *Liber Abbaci*, beginning from part-of-speech (PoS) tagging and lemmatization of a specific chapter of the book, chosen for its linguistic and historical interest. The dataset is freely available online[1].

This paper describes the process of annotation of the *Liber Abbaci* and two applications of its

[1] http://dialogo.di.unipi.it/LiberAbbaci

results, namely (a) the evaluation of a number of trained models for PoS tagging and lemmatization for Latin in out-of-domain fashion and (b) the interlinking of the annotated chapter with other linguistic resources for Latin through the Lila Knowledge Base (KB)[2].

2 Related Work

The research area dealing with the creation of linguistic resources and Natural Language Processing (NLP) tools for ancient languages has seen a remarkable growth during the last decade (Sprugnoli and Passarotti, 2020). This has primarily concerned Latin and Ancient Greek as essential media to access and understand the so-called Classical heritage. In particular, several annotated corpora of Latin texts are currently available in digital format: they follow different guidelines and tagsets and feature different layers of linguistic annotation. This section wants to provide a (far from exhaustive) overview of such resources to show how the dataset presented in this paper stands with respect to the state of the art.

The LASLA corpus contains 2,500,000 semi-manually annotated tokens. It covers a large portion of the extant Classical Latin literature. It was started in 1961 by the LASLA research center at the Université de Liège[3] and is still being expanded[4]. The corpus is considered to be a gold standard, since the annotation of every token has been manually verified by a philologist. The linguistic information consists of lemmatization, morphological tagging, and an additional syntactic layer for verbs (Verkerk et al., 2020). Texts cover various literary genres (theater, poetry, prose) and have a chronological extension ranging from the comedies of Plautus to the texts of Suetonius and Pliny the Younger. Recent additions reach later stages of Latin literature [5], but include neither Medieval nor Neo-Latin works. Natural sciences and technical works are weakly represented in the corpus, the treatise *De Agri Cultura* 'on agriculture' by Cato and the recently added *Naturales Quaestiones* 'investigations about nature' by Seneca being the only examples.

The corpus of Latin Lemmatized Texts released by Thibault Clérice (Clérice, 2021a) is formed by 21,222,911 tokens (17,804,769 without punctuation marks) and includes a large set of Classical and Late Latin texts available in a a number of open access corpora[6]. Clérice's corpus covers a very ample chronological span (up until the 9th century) as well as different genres: from Classical literature (Horace, Ovid, etc.), to Christian religious texts and legal texts. The linguistic annotation consists of lemmatization and full morphological description of the tokens , produced automatically by applying the Pie Latin LASLA+ model 0.0.6 (Manjavacas et al., 2019), fine-tuned on ca. 1,500,000 tokens taken from the LASLA corpus (Clérice, 2021b), with very good results concerning lemmatization and PoS tagging[7]. However, results appear to be less good on unknown tokens[8]. This difference underlines the difficulty of using automatic annotation tools on texts with a very specialized language, surely not found in LASLA, as is the case for Fibonacci's *Liber Abbaci*.

As for syntactically annotated corpora, five treebanks are currently available for Latin. They are the *Index Thomisticus* Treebank (IT-TB) (Passarotti, 2019), the PROIEL treebank (Haug and Jøhndal, 2008; Eckhoff et al., 2018), the Latin Dependency Treebank by the Perseus Digital Library (part of the Ancient Greek and Latin Treebank) (Bamman and Crane, 2007), the Late Latin Charter Treebank (LLCT) (Cecchini et al., 2020b) and the UDante treebank (Cecchini et al., 2020a). The treebanks include texts of different genres (literary, historical, philosophical and documentary) and periods (from Classical to Medieval), but technical works are not represented.

3 Dataset Creation and Analysis

The *Liber Abbaci* is made up of more than 270,000 tokens and is divided into 15 chapters of varying length. The choice of starting our manual annotation from chapter VIII *de reperiendis pretiis mercium per maiorem guisam* 'on finding out the price of goods through the "greater means"' is due to the

[2] https://lila-erc.eu

[3] http://web.philo.ulg.ac.be/lasla/presentation-du-laboratoire/

[4] See http://web.philo.ulg.ac.be/lasla/textes-latins-traites/.

[5] Of which some are already available: see http://web.philo.ulg.ac.be/lasla/textes-latins-en-cours-de-traitement/.

[6] For the full list, see https://github.com/lascivaroma/latin-lemmatized-texts/tree/0.1.2.

[7] For lemmatization, accuracy: 0.9734 . For PoS tagging, accuracy: 0.9651 .

[8] For lemmatization, accuracy: 0.8716 . For PoS tagging, accuracy: 0.9232 .

peculiarity of its content. Here, Fibonacci treats many simple business negotiations using proportions and referring to many examples taken from the entire Mediterranean world. The examples concern weight and monetary systems as well as the main products bought and sold in the 13th century. This means that the text is rich of terminology specific of the mathematical domain but also of trade and commerce. Chapter VIII is made up of 29,858 tokens (including punctuation marks), thus covering about 10% of the total length of the *Liber Abbaci*.

3.1 Data Annotation

The manual annotation of chapter VIII is carried out by a master's degree student in Classical languages, with excellent knowledge of Latin but without any previous expertise in either linguistic annotation or computational linguistics. The overall effort of the work amounts to a total of 227 hours, including: training sessions, study of the guidelines and of terminology related to measures, coins and trade in the Middle Ages (Marcinkowski, 2003; Martinori, 1915), the actual annotation, the reconciliation after evaluation of inter-annotator agreement (IAA, see Section 3.2), periodic checks with supervisors, the linking of the annotated text to the LiLa KB (see Section 5). We make use of a large number of dictionaries as references: the *Oxford Latin Dictionary* (OLD) (Souter, 1968), the *Lexicon Totius Latinitatis* (Forcellini, 1965), the *Dictionnaire illustré latin-français* (herafter: Gaffiot) (Gaffiot, 2016) and the *Thesaurus Linguae Latinae*[9] for Classical Latin, but also the *Dictionary of Medieval Latin from British Sources* (Latham and Howlett, 1975) and the *Glossarium mediae et infimae latinitatis* (du Cange et al., from 1883 to 1887) for Medieval Latin. Tokenization and sentence splitting are performed manually on a text editor, then lemmatization and PoS tagging are carried out on a shared spreadsheet following the Universal Dependencies (UD) formalism (de Marneffe et al., 2021), in particular both the universal and the language-specific guidelines relative to the latest release of the UD treebanks (v 2.9)[10].

The implementation of the UD guidelines to the linguistic peculiarities of the text does not always happen straightforwardly. Chapter VIII of the *Liber Abbaci*, as well as the work in its entirety, presents several typical features of Medieval Latin, both graphically (e. g. the monophthongization *ae* → *e* and the spelling *nichil* instead of the Classical *nihil* 'nothing'), morphologically (e. g the presence of analytical verb forms such as the "perfect", i. e. present perfective, subjunctive *habeat . . . honeratum*, instead of the Classical *onerauerit*, from *onero* 'to load') and syntactically (e. g. the nearly exclusive use of *quod* 'that' to introduce declarative clauses, instead of accusative and infinitive[11]). It is also worth noting the very limited use of enclitic particles (in the whole chapter VIII, Fibonacci uses the enclitic conjunction *que* 'and' only 3 times, appended to the auxiliary verb form *erunt* 'they will be') and the presence of syntactic calques of vernacular constructions (e. g. *secundum quod uadis multiplicando* 'according to what you are multiplying', where *uado* is preferred to the more Classical *eo* 'to go' and further assumes an auxiliary function, and the use of the gerundive form *multiplicando* is an innovation).

But the main peculiarities of the text concern the lexicon. Chapter VIII presents indeed a rich set of toponyms, units of measurement, names of coins and Arabisms often not even reported by Medieval Latin dictionaries. This is the case, for example, of some names of places, such as *Bugea*, today's Biğāya/Bgayet in Algeria (a city where Fibonacci spent a period of his childhood, learning the art of calculation), and *Septis*, today's Ceuta/Sabta on the Strait of Gibraltar; or, among the numismatic terms, of *bolsonalia*, a word designating a certain amount of broken silver or mixture coins which were sold to goldsmiths because they were adulterated or out of date.

3.2 Inter-Annotator Agreement

The IAA is calculated on 30 sentences (1,010 tokens), with the participation of a second scholar with a background in Classical languages. We register an almost perfect agreement with a Cohen's κ (Artstein and Poesio, 2008) of 0.97 for lemmatization and 0.94 for PoS tagging.

The comparison between the two annotations highlights two main issues. The first concerns the choice of the UPOS (Universal Part Of Speech) tag (de Marneffe et al., 2021, §2.2.2) for terms such as

[9] https://thesaurus.badw.de/das-projekt.html

[10] https://universaldependencies.org/guidelines.html

[11] See for example (Traina and Bertotti, 2015, C. XVI) .

nam 'certainly' and *enim* 'namely', because different corpora and dictionaries adopt different conventions: e. g. *nam* is labeled as `adverb` in the Lila KB and `Df` in the Latin PROIEL treebank, both possibly equivalent to UPOS `ADV`[12]; as `S`[13], standing for *conjonction de coordination* (UPOS: `CCONJ`) in the LASLA corpus, and more generically *conjonction (servant à confirmer/causale)* (UPOS either `CCONJ` or `SCONJ`) in the Gaffiot; finally *particle* (not necessarily corresponding to UPOS `PART`) in the OLD, and similarly *particule* in one sense in the Gaffiot. The treatment of the etymologically related and functionally similar *enim* is mostly identical for all sources, only with the Gaffiot reporting a sense as *adverbe* instead of *particule*, followed by the LASLA corpus in using both labels `S` and `M` (generic for *adverbe*), the latter though very marginally. These terms have been discussed and finally assigned the UPOS `PART`, used in the latest Latin UD treebanks to label discoursive particles like these. Such difficulties derive on one hand from the "volatile" and diachronically variable nature of similar elements, but on the other hand, and relatedly, to traditional grammars overlooking them and more generally skipping over pragmatic phenomena, in favour of "more Classical" parts of speech (hence the frequent inclusion of *nam*, *enim*, etc. in the catchall category of "adverbs").

The second issue is the UPOS to be used for *unus* 'one'. Fibonacci often uses *unus* to indicate a generic entity, as is clearly visible when paralleled by *alter* 'other'. In this case, *unus* is tagged as `DET` (determiner), like *alter*[14]. In a number of other contexts, however, *unus* specifies the quantity of a certain object. In such cases it is considered a `NUM` (numeral)[15]. The difficulty here originates from a well known and complex linguistic change that will eventually produce a clear indefinite article from the numeral in Romance languages, but for which, being so gradual, we cannot pinpoint an exact historical moment; cf. (Ledgeway, 2012, §4.2.1).

4 Comparing NLP Models

Table 1 reports accuracy scores computed on our gold standard processed with UDPipe using the UD v2.6 models for Latin (Straka and Straková, 2017). The scores clearly show that current models are not good enough to process the Latin of Fibonacci. The best accuracy for lemmatization is achieved by the model trained on the LLCT treebank, which contains a set of Early Medieval charters written in Tuscany. However, this scores are lower than state-of-the-art ones: the best participating system at the EvaLatin 2020 evaluation campaign achieves an accuracy of $96,19\%$ for lemmatization and $96,74\%$ for PoS tagging on the corresponding test set (Sprugnoli et al., 2020), i. e. about 33 and 15 points more than the results obtained on Fibonacci.

	Lemma	UPOS
EvaLatin2020	63.60	81.90
IT-TB	65.58	77.14
LLCT	68.81	82.79
Perseus	67.54	78.37
PROIEL	60.25	51.64

Table 1: Accuracy of UDPipe v2.6 Latin models tested on chapter VIII of the *Liber Abbaci*.

Taking into consideration lemmatization, the percentage of out-of-vocabulary lemmas, that is, lemmas present in the text by Fibonacci but not in the training texts of the models, is very high ($> 50\%$ of lemma types). The majority of errors are registered for numbers and common nouns. The first problem is due to the fact that some models do not recognize Arabic numbers, because they have not seen them in their training data, while others lemmatize them with a special "metalemma" of the kind of *num. arab.*, eschewing lexical forms. As for common nouns, most errors related to lemmatization concern the lexical classes discussed in Section 3. For example, the tokens *libris* and *libre* are often lemmatized as *liber* 'free' (`ADJ`) instead of *libra* 'pound' (`NOUN`).

Table 2 shows the F1 score per UPOS tag. We observe that an F1 above 70% is achieved by any model only on 5 tags: `ADP`, `NOUN`, `NUM`, `SCONJ` and `VERB`. No model recognizes the `SYM` tag (used for mathematical operators such as paren-

[12] Cf. (Eckhoff et al., 2018, §5)

[13] With only very few exceptions when it is seen as part of a compound expression with tmesis, thus not receiving an autonomous PoS; cf. Pl. Am. 2.1, 49-50: ***Quo** id, malum, pacto potest **nam** (mecum argumentis puta) fieri, nunc uti tu et hic sis et domi?*, interpreted as an instance of *quonam* 'whither pray?', itself receiving `K` meaning *pronom interrogatif.*

[14] For instance, in the clause *ita est pretium **unius** ad pretium alterius* (VIII, 8) 'so the price of **the one** [merchandise] is to the price of the other'.

[15] For instance, in the clause … *que multiplica per summam denariorum **unius** libre* (VIII, 20) 'which you have to multiply by the amount of *denarii* of which **one** pound consists'.

	EvaLatin2020	IT-TB	Perseus	PROIEL	LLCT
SYM	0.00	0.00	0.00	0.00	0.00
AUX	0.24	0.45	0.03	0.27	0.32
ADJ	0.48	0.37	0.40	0.28	0.55
PRON	0.57	0.38	0.40	0.52	0.93
PART	0.65	0.00	0.00	0.00	0.65
ADV	0.70	0.71	0.78	0.21	0.84
CCONJ	0.75	0.67	0.68	0.44	0.86
SCONJ	0.89	0.95	0.96	0.86	0.95
VERB	0.91	0.87	0.92	0.78	0.84
NOUN	0.92	0.83	0.86	0.75	0.88
DET	0.93	0.00	0.00	0.53	0.91
PROPN	0.94	0.09	0.00	0.32	0.57
NUM	0.95	0.96	0.96	0.75	0.99
ADP	0.99	0.98	0.93	0.91	0.88
Global	0.71	0.52	0.49	0.46	0.73

Table 2: F1 on UPOS tags of UDPipe v2.6 Latin models on chapter VIII of the *Liber Abbaci*.

theses), because it is not present in their respective training data. The same is true for the tag PART in IT-TB (up until UD v2.8)[16], Perseus and PROIEL, and for the tag DET in Perseus. In old versions of the IT-TB, DET is limited to the proto-article *ly* (8 occurrences), while in Perseus the tag PROPN appears only for the lemma *Aefulanus* (1 occurrence). The IT-TB-based model, too, registers a near-zero F1 score for PROPN: in the corresponding training data, this tag is used for a restricted (116 types of lemmas) set of terms mostly specific to the domains of philosophy and religion (e. g. *Aristoteles*, *Maria*), not present in our dataset. Low performances are registered also for the AUX tag, the annotation of which is not consistent in training data: in Perseus, this tag is not used at all, while in EvaLatin 2020 it marks only the auxiliaries in periphrastic passive (including deponent) constructions, while in the other treebanks it is applied also to verbal copulas, as per UD guidelines. Further, the *Liber Abbaci* sees the rise (1 occurrence) of *habeo* 'to have' as a possible auxiliary (cf. Section 3.1), unheard of in Classical Latin and only attested (albeit marginally) in LLCT.

5 Linking and Querying in LiLa

The LiLa KB makes linguistic resources for Latin interoperable by linking tokens in corpora and entries in dictionaries/lexica to a collection of canonical forms for Latin called Lemma Bank (Passarotti et al., 2020). In order to connect the lemmas of chapter VIII to LiLa's KB, a string match is first performed between the lemmas in the texts and those in the KB, also taking into account their parts of speech. Using this strategy, 88.8% of the lemmas are directly connected to a single entry in the KB. The remaining unconnected lemmas fall into two possible categories: ambiguous lemmas, that is, with possible connection to more than one entry in the KB; and lemmas absent from the KB. More specifically, we find 44 ambiguous lemmas (corresponding to 631 tokens): for example, *colligo* can be connected to two entries: either a first-conjugation verb *colligare*[17] 'to bind', or a third-conjugation verb *colligĕre*[18] 'to gather'. These cases are manually disambiguated, checking each context of use. The remaining, not directly connected lemmas are not present in the KB and need to be manually added: these are mainly words denoting weight and monetary units (e. g. *karatus* 'carat'), or different written representations of lemmas already in LiLa (e. g. *torscellus* is a graphic variant of *tor-*

[16] Annotation discrepancies with respect to other Latin UD treebanks for INTJ, NUM, PART, PRON and DET have been resolved in IT-TB in its last version (2.9), released in November 2021; however, the model adopted in this paper and currently available in UDPipe is based on an older version of the data.

[17] https://lila-erc.eu/data/id/lemma/94854

[18] https://lila-erc.eu/data/id/lemma/94855

cellus[19], a unit of length). Thanks to the linking, each lemma of our dataset becomes part of an interoperable ecosystem made of resources of different kinds. We can thus query different interlinked resources using SPARQL and the LiLa endpoints[20]. For example, we can find the lemmas appearing only in chapter VIII[21] and not in the other texts that are currently linked to the KB: the *Summa Contra Gentiles* by Thomas Aquinas (from the *Index Thomisticus*), those found in UDante (a corpus of 5 works mostly by Dante Alighieri, or attributed to him, manually annotated following the UD formalism), and the *Querolus siue Aulularia* (an anonymous comedy dating back to the 5th c. AD).

Lemma	Gloss	Freq.
rotulus	unit of weight	296
soldus	monetary unit	212
virgula	bar of a fraction	202
byzantius	monetary unit	73
cantare	unit of weight	67

Table 3: The 5 most frequent distinctive lemmas in chapter VIII of the *Liber Abbaci*.

Table 3 shows the 5 most frequent distinctive, i. e. exclusively found in the *Liber Abbaci*, lemmas retrieved using a SPARQL query[22]. They are all related to mathematics, coins and units of measurement, confirming the specificity of the domain of our dataset. In particular, *rotulus* and *cantāre* are two units of weight, both deriving from Arabic, respectively from *raṭl* (in turn, a metathetical adaptation of Greek λίτρα *litra* 'pound') and *qinṭār*, which designates a weight of 100 *rotuli*[23]. The term *soldus*, instead, indicates a unit of measurement used for monetary quantities. Among the many currencies mentioned in chapter VIII, Fibonacci often cites the *byzantius*, a golden coin minted in Constantinople[24]. Finally, *virgula* (diminutive of *virga*, properly a 'rod', used by Fibonacci in the same sense of *virgula*) primarily denotes the bar between the numerator and denominator of a fraction, but it can also designate the fraction itself (Bocchi, 2004).

6 Conclusions and Future Work

This paper describes the annotation of one chapter of the *Liber Abbaci* by Fibonacci, and reports on the linguistic peculiarities of this text and the ensuing challenges.

The results of existing UDPipe models in lemmatization and tagging show low accuracy and F1 scores when compared to the state of the art for these tasks in the recent EvaLatin 2020 evaluation campaign. This, on the one hand, can be attributed to the characteristics of the genre of Fibonacci's texts, which are representative of scientific Medieval Latin texts, and on the other hand can be explained with the different choices in annotation style of Latin treebanks released under the UD project. Substantial improvements can be expected with models trained on new releases of Latin treebanks which have already undertaken the effort of resolving annotation discrepancies and of making the annotation style across treebanks more homogeneous. Further improvements will however require new annotated chapters and experiments in domain adaptation, which are scheduled as future work.

Acknowledgments

This work is a contribution to the *Fibonacci 1202-2021* project, financed by the Tuscany Region. Part of the work has been funded by the European Research Council (ERC) under the European Union's *Horizon 2020* research and innovation programme – Grant Agreement No. 769994. The authors want to thank: prof. Andrea Bocchi, dott. Alessandro Gelsumini, prof. Pier Daniele Napolitani and prof. Enrica Salvatori for their linguistic and historical advice.

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

[19] https://lila-erc.eu/data/id/lemma/133810

[20] https://lila-erc.eu/sparql/

[21] https://lila-erc.eu/data/corpora/CorpusFibonacci/id/corpus/Liber Abbaci

[22] https://github.com/CIRCSE/SPARQL-queries/blob/main/distinctivelemmas-Fibonacci.rq

[23] It should be noted that Fibonacci alternates a third-declension *cantāre* (gen. sing. *cantāris*) with a second-declension *cantarium* (gen. sing. *cantarii*). During lemmatization of the text, the various attested singular forms have been linked to their respective lemmas; the nom./acc. plur. *cantaria*, which theoretically could derive both from *cantāre* and *cantarium*, has been linked to the lemma *cantāre* for simple reasons of probability, as it is the most frequently used by Fibonacci among these two forms.

[24] Also mentioned is the *byzantius saracenatus*, equivalent to the *hyperperus*, that is, a *byzantius* with inscriptions in Kufic characters (Martinori, 1915).

David Bamman and Gregory Crane. 2007. The Latin Dependency Treebank in a Cultural Heritage Digital Library. In *Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007)*, pages 33–40, Prague, Czech Republic, June. Association for Computational Linguistics.

Andrea Bocchi. 2004. In Michelangelo Zaccarello and Lorenzo Tomasin, editors, *Storia della lingua e filologia. Per Alfredo Stussi nel suo sessantacinquesimo compleanno*, chapter Sì nel Livero de l'abbecho, pages 121–158. SISMEL – Edizioni del Galluzzo, Florence, Italy.

Flavio M. Cecchini, Rachele Sprugnoli, Giovanni Moretti, and Marco Passarotti. 2020a. UDante: First Steps Towards the Universal Dependencies Treebank of Dante's Latin Works. In *Seventh Italian Conference on Computational Linguistics*, pages 1–7, Bologna. CEUR-WS.org.

Flavio Massimiliano Cecchini, Timo Korkiakangas, and Marco Passarotti. 2020b. A New Latin Treebank for Universal Dependencies: Charters between Ancient Latin and Romance Languages. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 933–942, Marseille, France, May. European Language Resources Association.

Thibault Clérice. 2021a. lascivaroma/latin-lemmatized-texts: 0.1.2 - HN PSL, May. DOI: `10.5281/zenodo.4661034`; project online at `https://github.com/lascivaroma/latin-lemmatized-texts`.

Thibault Clérice. 2021b. Latin Lasla Model, Apr. DOI: `10.5281/zenodo.4661034`.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308, 07.

Charles du Fresne sieur du Cange, bénédictins de la congrégation de Saint-Maur, d. Pierre Carpentier, Johann Christoph Adelung, G. A. Louis Henschel, Lorenz Diefenbach, and Léopold Favre. from 1883 to 1887. *Glossarium mediae et infimae latinitatis*. Favre, Niort, France.

Hanne Martine Eckhoff, Kristin Bech, Gerlof Bouma, Kristine Eide, Dag Haug, Odd Einar Haugen, and Marius Jøhndal. 2018. The PROIEL treebank family: a standard for early attestations of Indo-European languages. *Language Resources and Evaluation*, 52(1):29–65.

Leonardus Bigollus Pisanus vulgo Fibonacci. 2020. *Liber Abbaci*, volume 79 of *Biblioteca di «Nuncius»*. Leo S. Olschki, Florence, Italy.

Egidio Forcellini. 1965. *Lexicon totius latinitatis*. Arnaldo Forni, Bologna, Italy.

Félix Gaffiot. 2016. *Dictionnaire Latin-Français*. Accessible at `gaffiot.fr`.

Dag Trygve Truslew Haug and Marius Jøhndal. 2008. Creating a Parallel Treebank of the Old Indo-European Bible Translations. In *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, pages 27–34.

Ronald Edward Latham and David R Howlett. 1975. *Dictionary of Medieval Latin from British Sources: Fascicule V: IJKL*. OUP Oxford.

Adam Ledgeway. 2012. *From Latin to Romance*, volume 1 of *Oxford studies in historical and diachronic linguistics*. Oxford University Press, Oxford, UK.

Enrique Manjavacas, Ákos Kádár, and Mike Kestemont. 2019. Improving lemmatization of non-standard languages with joint learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1493–1503, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Hinz Marcinkowski. 2003. *Measures and Weights in the Islamic World, an English Translation of Walther Hinz's Handbook Islamische Maße und Gewichte*. International Islamic University Malaysia (IIUM).

Edoardo Martinori. 1915. *La Moneta: vocabolario generale*. Instituto italiano di numismatica.

Marco Passarotti, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo, and Rachele Sprugnoli. 2020. Interlinking through lemmas. the lexical collection of the lila knowledge base of linguistic resources for latin. *Studi e Saggi Linguistici*, 58(1):177–212.

Marco Passarotti, 2019. volume 10 of *Age of Access? Grundfragen der Informationsgesellschaft*, chapter The Project of the Index Thomisticus Treebank, pages 299–320. De Gruyter Saur, Berlin, Germany; Boston, MA, USA.

Alexander Souter. 1968. *Oxford Latin dictionary: OLD*. Clarendon Press.

Rachele Sprugnoli and Marco Passarotti. 2020. Proceedings of LT4HALA 2020-1st Workshop on Language Technologies for Historical and Ancient Languages. In *Proceedings of LT4HALA 2020-1st Workshop on Language Technologies for Historical and Ancient Languages*.

Rachele Sprugnoli, Marco Passarotti, Flavio Massimiliano Cecchini, and Matteo Pellegrini. 2020. Overview of the EvaLatin 2020 evaluation campaign. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 105–110, Marseille, France, May. European Language Resources Association (ELRA).

Milan Straka and Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August. Association for Computational Linguistics.

Alfonso Traina and Tullio Bertotti. 2015. *Sintassi normativa della lingua latina*. Pàtron, Bologna, Italy.

Philippe Verkerk, Yves Ouvrard, Margherita Fantoli, and Dominique Longrée. 2020. L.A.S.L.A. and Collatinus: a convergence in lexica. *SSL*, 1(LVIII):95–120.