Vir is to *Moderatus* as *Mulier* is to *Intemperans* Lemma Embeddings for Latin

Rachele Sprugnoli, Marco Passarotti, Giovanni Moretti
CIRCSE Research Centre, Università Cattolica del Sacro Cuore
Largo Agostino Gemelli 1, 20123 Milano
{rachele.sprugnoli,marco.passarotti,giovanni.moretti}@unicatt.it

Abstract

English. This paper presents a new set of lemma embeddings for the Latin language. Embeddings are trained on a manually annotated corpus of texts belonging to the Classical era: different models, architectures and dimensions are tested and evaluated using a novel benchmark for the synonym selection task. A qualitative evaluation is also performed on the embeddings of rare lemmas. In addition, we release vectors pre-trained on the "Opera Maiora" by Thomas Aquinas, thus providing a resource to analyze Latin in a diachronic perspective.[1]

1 Introduction

Any study of the ancient world is inextricably bound to empirical sources, be those archaeological relics, artifacts or texts. Most ancient texts are written in dead languages, one of the distinguishing features of which is that both their lexicon and their textual evidence are essentially closed, without any new substantial addition. This finite nature of dead languages, together with the need of empirical data to their study, makes the preservation and the careful analysis of their legacy a core task of the (scientific) community. Although computational and corpus linguistics have mainly focused on building tools and resources for modern languages, there has always been large interest in providing scholars with collections of texts written in dead or historical languages (Berti, 2019). Not by chance, one of the first electronic corpora ever produced is the "Index Thomisticus" (Busa, 1974 1980), the opera omnia of Thomas Aquinas written in Latin in the 13th century. Owing to its wide diachronic span covering more than two millennia, as well as its diatopic distribution across Europe and the Mediterranean, Latin is the most resourced historical language with respect to the availability of textual corpora. Large collections of Latin texts, e.g. the *Perseus Digital Library*[2] and the corpus of Medieval Italian Latinity *ALIM*[3], can now be processed with state-of-the-art computational tools and methods to provide linguistic resources that enable scholars to exploit the empirical evidence provided by such datasets to the fullest. This is particularly promising given that the quality of many textual resources for Latin, carefully built over decades, is high.

Recent years have seen the rise of language modeling and feature learning techniques applied to linguistic data, resulting in so-called "word embeddings", i.e. empirically trained vectors of lexical items in which words occurring in similar linguistic contexts are assigned close vectorial space. The semantic meaningfulness and motivation of word embeddings stems from the basic assumption of distributional semantics, according to which the distributional properties of words mirror their semantic similarities and/or differences, so that words sharing similar contexts tend to have similar meanings.

In this paper, we present and evaluate a number of embeddings for Latin built from a manually lemmatized dataset containing texts from the Classical era.[4] In addition, we release embeddings trained on a manually lemmatized corpus of medieval texts to facilitate diachronic analyses. This research is performed in the context of the *LiLa: Linking Latin* project, which seeks to build a Knowledge Base of linguistic resources for Latin connected via a common vocabulary of knowledge

[2] http://www.perseus.tufts.edu/hopper/

[3] http://www.alim.dfll.univr.it/

[4] Word embeddings built on tokens of the same dataset are also available online.

description following the principles of the Linked Data framework.[5] Our contribution provides the community with new resources to be connected in the LiLa Knowledge Base aimed at supporting data-driven socio-cultural studies of the Latin world. The added value of our lemma embeddings for Latin results from the interdisciplinary blending of state-of-the-art methods in computational linguistics with the long tradition of Latin corpora creation: on the one hand the embeddings are evaluated with techniques hitherto applied to modern languages data only, on the other they are built from high quality datasets heavily used by scholars working on Latin.

2 Related Work

Word embeddings are crucial to many Natural Language Processing (NLP) tasks (Collobert et al., 2011; Lample et al., 2016; Yu et al., 2017). Numerous pre-trained word vectors generated with different algorithms have been released, typically generated from huge amounts of contemporary texts written in modern languages. The interest towards this type of distributional approach has emerged also in the Digital Humanities, as evidenced by publications on the use of word embeddings trained on literary texts or historical documents (Hamilton et al., 2016; Leavy et al., 2018; Sprugnoli and Tonelli, 2019). Although to a lesser extent, the literature also reports works on word embeddings for dead languages, including Latin.

Both Facebook and the organizers of the CoNLL shared tasks on multilingual parsing have pre-computed and released word embeddings trained on Latin texts crawled from the web: the former using the fastText model on Common Crawl and Wikipedia dumps (Grave et al., 2018a), the latter applying word2vec to Common Crawl only (Zeman et al., 2018). Both resources were developed by relying on automatic language detection engines: they are very big in terms of vocabulary size[6] but highly noisy due to the presence of languages other than Latin. In addition, they include terms related to modern times, such as movie stars, TV series, companies (e.g., *Cumberbatch*, *Simpson*, *Google*), making them unsuitable for the study of language use in ancient texts. The automatic detection of language has also been employed by Bamman (2012) to collect a corpus of Latin books available from Internet Archive. The corpus spans from 200 BCE to the 20th century and contains 1.38 billion tokens: embeddings trained on this corpus[7] were used to investigate the relationship between concepts and historical characters in the work of Cassiodorus (Bjerva and Praet, 2015). However, these word vectors are affected by OCR errors present in the training corpus: 25% of the embedding vocabulary contains non-alphanumeric characters, e.g. -**-, *iftud^*. The quality of the corpus used to train the Latin word embeddings available through the SemioGraph interface[8], on the other hand, is high: these embeddings are based on the "Computational Historical Semantics" database, a manually curated collection of 4,000 Latin texts written between the 2nd and the 15th century AD (Jussen and Rohmann, 2015). In SemioGraph, more than one hundred word vectors can be visually explored searching by Part-of-Speech (PoS) labels and text genres: however, these vectors cannot be downloaded for further analysis and were generated with one model only, i.e. word2vec.

With respect to the works cited above, in this paper we rely on manually lemmatized texts free of OCR errors, we focus on a period not covered by the "Computational Historical Semantics" database and we test two models to learn lemma representations. It is worth noting that none of the previously mentioned studies have carried out an evaluation of the trained Latin embeddings; we, on the contrary, provide both quantitative and qualitative evaluations of our vectors.

3 Dataset Description

Our lemma vectors were trained on the "Opera Latina" corpus (Denooz, 2004). This textual resource has been collected and manually annotated since 1961 by the Laboratoire d'Analyse Statistique des Langues Anciennes (LASLA) at the University of Liège[9]. It includes 158 texts from 20 different Classical authors covering various genres, such as treatises (e.g. "Annales" by Tacitus), letters (e.g. "Epistulae" by Pliny the Younger), epic poems (e.g. "Aeneis" by Virgil), elegies

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

[6] For example, the size of the CoNLL embeddings vocabulary is 1,082,365 words.

[7] http://www.cs.cmu.edu/~dbamman/latin.html

[8] http://semiograph.texttechnologylab.org/

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

TARGET WORDS	SYNONYMS	DECOY WORDS		
decretum/decree	*edictum*/proclamation	*flagitium*/shameful act	*adolesco*/to grow up	*stipendiarius*/tributary
saepe/often	*crebro*/frequently	*conquiro*/to seek for	*ululatus*/howling	*frugifer*/fertile
rogo/to ask	*oro*/to ask for	*columna*/column	*retorqueo*/to twist back	*errabundus*/vagrant
exilis/thin	*macer*/emaciated	*moles*/pile	*mortalitas*/mortality	*audens*/daring

Table 1: Examples taken from the Latin benchmark for the synonym selection task.

(e.g. “Elegiae” by Propertius), plays (both comedies and tragedies e.g. “Aulularia” by Plautus and “Oedipus” by Seneca), and public speeches (e.g. “Philippicae” by Cicero)[10].

The corpus contains several layers of linguistic annotation, such as lemmatization, PoS tagging and tagging of inflectional features, organized in space-separated files. “Opera Latina” contains approximately 1,700,000 words (punctuation is not present in the corpus), corresponding to 133,886 unique tokens and 24,339 unique lemmas.

4 Experimental Setup

We tested two different vector representations, namely word2vec (Mikolov et al., 2013a) and fastText (Bojanowski et al., 2017): the former is based on linear bag-of-words contexts generating a distinct vector for each word, whereas the latter is based on a bag of character n-grams, that is, the vector for a word (or a lemma) is the sum of its character n-gram vectors. Lemma vectors were pre-computed using two dimensionalities (100, 300) and two models: skip-gram and Continuous Bag-of-Words (CBOW). In this way, we had the possibility of evaluating both modest and high dimensional vectors and two architectures: skip-gram is designed to predict the context given a target word, whereas CBWO predicts the target word based on the context. The window size was 10 lemmas for skip-gram and 5 for CBOW. The other training options were the same for the two models:

- number of negatives sampled: 25;
- number of threads: 20;
- number of iterations over the corpus: 15;
- minimal number of word occurrences: 5.

Embeddings were trained on the lemmatized “Opera Latina” in order to reduce the data sparsity due to the high inflectional nature of Latin. Moreover, we lower-cased the text and converted *v* into *u* (so that *vir* ‘man’ becomes *uir*) to fit the lexicographic conventions of some Latin dictionaries (Glare, 1982) and corpora. With the minimal number of lemma occurrences set to 5, we obtained a vocabulary size of 11,327 lemmas.

	word2vec		**fastText**	
	cbow	skip-gram	cbow	skip-gram
100	81.14%	79.83%	80.57%	**86.91%**
300	80.86%	79.48%	79.43%	86.40%

Table 2: Results of the synonym selection task calculated on the whole benchmark.

	word2vec		**fastText**	
	cbow	skip-gram	cbow	skip-gram
100	81.48%	85.18%	77.77%	87.03%
300	76.63%	85.18%	75.92%	**90.74%**

Table 3: Results of the synonym selection task calculated on a subset of the benchmark containing only questions with lemmas sharing the same PoS.

5 Evaluation

Word embeddings resulting from the experiments described in the previous Section were tested performing both an intrinsic and a qualitative evaluation (Schnabel et al., 2015). To the best of our knowledge, these methods, although well documented in the literature, have never been applied to the evaluation of Latin embeddings.

5.1 Synonym Selection Task

In the synonym selection task, the goal is to select the correct synonym of a target lemma out of a set of possible answers (Baroni et al., 2014). The most commonly used benchmark for this task is the Test of English as a Foreign Language (TOEFL), consisting of multiple-choice questions each involving five terms: the target words and another four, one of which is a synonym of the target word and the remaining three decoys (Landauer and Dumais, 1997). The original TOEFL dataset is made of only 80 questions but extensions have been proposed to widen the set of multiple-choice questions using external resources such as WordNet (Ehlert, 2003; Freitag et al., 2005).

In order to create a TOEFL-like benchmark for Latin, we relied on four digitized dictionaries

[10] The corpus can be queried through an online interface after requesting credentials: `http://cipl93.philo.ulg.ac.be/OperaLatina/`

	contrudo/to thrust	**frugaliter**/thriftily	**auspicatus**/consecrated by auspices
fastText-skip	*protrudo**/to thrust forward *extrudo**/to thrust out	*frugalis**/thrifty *frugalitas**/economy	*auspicato**/after taking the auspices *auspicium**/auspices
fastText-cbow	*contego**/to cover *contraho*/to collect	*aliter*/differently *negligenter*/neglectfully	*auguratus**/the office of augur *pontificatus*/the office of pontifex
word2vec-skip	*infodio*/to bury *tabeo*/to melt away	*frugi**/frugal *quaerito*/to seek earnestly	*erycinus*/Erycinian *parilia*/the feast of Pales
word2vec-cbow	*refundo*/to pour back *infodio*/to bury	*lautus*/neat *frugi**/frugal	*erycinus*/Erycinian *parilia*/the feast of Pales

Table 4: Examples of the nearest neighbors of rare lemmas

of Latin synonyms (Hill, 1804; Dumesnil, 1819; Von Doederlein and Taylor, 1875; Skřivan, 1890) available online in XML Dictionary eXchange format[11]. Starting from the digital versions of the dictionaries, we proceeded as follows:

- we downloaded and parsed the XML files so as to extract only the information useful for our purposes, that is, the dictionary entry and the synonyms;
- we merged the content of all dictionaries to obtain the largest possible list of lemmas with their corresponding synonyms. Unlike "Opera Latina" and the other synonym dictionaries, Dumesnil (1819) often lemmatizes verbs under the infinite form; therefore, for the sake of uniformity, we used LEMLAT v3[12] to obtain the first person, singular, present, active (or passive, in case of deponent verbs), indicative form of all verbs registered in that dictionary in their present infinite form (e.g. *accingere* 'to gird on'→ *accingo*) (Passarotti et al., 2017). At the end of this phase, we obtained a new resource containing 2,759 unique entries and covering all types of PoS, together with their synonyms;
- multiple-choice questions were created by taking each entry as a target lemma, then adding its first synonym and another three lemmas randomly chosen from the "Opera Latina" corpus;
- a Latin language expert manually checked samples of multiple-choice questions so as to be sure that the three randomly chosen lemmas were in fact decoy lemmas.

Table 1 provides some examples of the multiple-choice questions generated using the procedure described above .

We computed the performance of the embeddings by calculating the cosine similarity between the vector of the target lemma and that of the other lemmas, picking the candidate with the largest cosine. Questions containing lemmas not included in the vocabulary, and thus vectorless, are automatically filtered out; results are given in terms of accuracy. As shown in Table 2, fastText proved to be the best lemma representation for the synonym selection task with the skip-gram architecture achieving an accuracy above 86%. This result can be explained by the fact that fastText is able to model morphology by taking into consideration sub-word units (i.e. character n-grams) and joining lemmas from the same derivational families. In addition, the skip-gram architecture works well with small amounts of training data like ours. It is also worth noting that, for both architectures and models, vectors with a modest dimensionality achieved a slightly higher accuracy with respect to embeddings with 300 dimensions.

The error analysis revealed specific types of linguistic and semantic relations, other than synonymy, holding between the target lemma and the decoy lemma that resulted having the largest cosine: for example, meronymy (e.g., target word: *annalis* 'chronicles' - synonym: *historia* 'narrative of past events' - answer: *charta* 'paper') and morphological derivation (e.g. target word: *consors* 'having a common lot' - synonym: *particeps* 'sharer' - answer: *sors* 'lot').

As an additional analysis, we repeated our evaluation on a subset of the benchmark containing 85 questions made of lemmas sharing the same PoS, e.g. *auxilior* 'to assist', *adiuuo* 'to help', *censeo* 'to assess', *reuerto* 'to turn back', *humo* 'to bury'. Results reported in Table 3 confirm that the skip-gram architecture provides the best accuracy for this task achieving a score above 90% for fastText embeddings with 300 dimensions. We also note an improvement of the accuracy for word2vec (+5%). The reasons behind these results need further in-

[11] `https://github.com/nikita-moor/latin-dictionary`

[12] `https://github.com/CIRCSE/LEMLAT3`

vestigations.

5.2 Qualitative Evaluation on Rare Lemma Embeddings

One of the main differences between word2vec and fastText is that the latter is supposed to be able to generate better embeddings for words that occur rarely in the training data. This is due to the fact that rare words in word2vec have few neighbor context words from which to learn the vector representation, whereas in fastText even rare words share their character n-grams with other words, making it possible to represent them reliably. To validate this hypothesis, we performed a qualitative evaluation of the nearest neighbors of a small set of randomly selected lemmas appearing between 5 and 10 times only in the "Opera Latina" corpus. Two Latin language experts manually checked the two most similar lemmas (in terms of cosine similarity) induced by the different 100-dimension embeddings we trained. Table 4 presents a sample of the selected rare lemmas and their neighbors: an asterisk marks neighbors that two experts judged as most semantically-related to the target lemma. This manual inspection, even if based on a small set of data, shows that the embeddings trained using the fastText model with the skip-gram architecture can find more similar lemmas that those trained with other models and architectures.

6 A Diachronic Perspective

Diachronic analyses are particularly relevant for Latin given that its use spans more than two millennia. To support this type of study we release, together with the embeddings presented in the previous Sections, lemma vectors trained on the "Opera Maiora", written by Thomas Aquinas in the 13th century. "Opera Maiora" is a set of philosophical and religious works comprising some 4.5 million words (Passarotti, 2015): all texts are manually lemmatized and tagged at the morphological level (Passarotti, 2010) and are part of the "Index Thomisticus" (IT) corpus.

Before training the embeddings, we preprocessed the texts following the conventions adopted in "Opera Latina": we lower-cased, removed punctuation, and converted *v* and *j* into *u* and *i*, respectively. Embeddings were trained with the configuration that reported the best results in the evaluation described in Section 5 (i.e. fastText with the skip-gram architecture and 100 dimensions). For a comparative analysis with the embeddings of "Opera Latina", we aligned the embeddings of "Opera Maiora" to the same coordinate axes using the unsupervised alignment algorithm provided with the fastText code (Grave et al., 2018b). Thanks to this alignment, we can inspect the nearest neighbors (nn) of lemmas in the two embeddings. For example, the lemma *ordo* shifts from social class or military rank (among the top 10 nn in the "Opera Latina" embeddings we find, in this order, *equester* 'cavalry', *legionarius* 'legionary', *turmatim* 'by squadrons') to referring to the concept of order and intellectual structure in Thomas Aquinas (nn in "Opera Maiora": *ordinatio* 'setting in order', *coordinatio* 'arranging together', *ordino* 'set in order') (Busa, 1977). Another interesting case is *spiritus*: in the Classical era it refers to 'breath' (nn in "Opera Latina": *spiro* 'to blow', *exspiro* 'to exhale', *spiramentum* 'draught'), while in Aquinas' Christian writings it associated with the Holy Ghost (nn: *sanctio* 'to make sacred', *donum* 'gift', *paracletus* 'protector') (Busa, 1983).

7 Conclusion and Future Work

In this paper we presented a new set of Latin embeddings based on high quality lemmatized corpora and a new benchmark for the synonym selection task. The aligned embeddings can be visually explored through a web interface and all the resources are freely available online: `https://embeddings.lila-erc.eu`.

Several future works are envisaged. For example, we plan to develop new benchmarks, like the analogy test (Mikolov et al., 2013b) or the rare words dataset (Luong et al., 2013), for the intrinsic quantitative evaluation of Latin embeddings. Moreover, embeddings could be used to improve the linking of datasets in the LiLa Knowledge Base. We would also like to extend the diachronic analysis to the embeddings trained on the "Computational Historical Semantics" database as soon as these become available.

This work represents the first step towards the development of a new set of resources for the analysis of Latin. This effort is laying the foundations of the first campaign devoted to the evaluation of NLP tools for Latin, EvaLatin.

Acknowledgments

This work is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme via the "LiLa: Linking Latin" project - Grant Agreement No. 769994. The authors also wish to thank Andrea Peverelli for his expert support on Latin and Chris Culy for providing his code for the embeddings visualization.

References

David Bamman and David Smith. 2012. Extracting two thousand years of Latin from a million book library. *Journal on Computing and Cultural Heritage (JOCCH)*, 5(1):1–13.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 238–247.

Monica Berti. 2019. *Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution*, volume 10. Walter de Gruyter GmbH & Co KG.

Johannes Bjerva and Raf Praet. 2015. Word embeddings pointing the way for late antiquity. In *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 53–57.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Roberto Busa. 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.

Roberto Busa. 1977. Ordo dans les oeuvres de st. thomas d'aquin. *II Coll. Intern. del Lessico Intellettuale Europeo*, pages 59–184.

Roberto Busa. 1983. De voce spiritus in operibus s. thomae aquinatis. *IV Coll. Intern. del Lessico Intellettuale Europeo*, pages 191–222.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(Aug):2493–2537.

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

Jean Baptiste Gardin Dumesnil. 1819. *Latin Synonyms: With Their Different Significations: and Examples Taken from the Best Latin Authors*. GB Whittaker.

Bret R Ehlert. 2003. *Making accurate lexical semantic similarity judgments using word-context cooccurrence statistics*. University of California, San Diego.

Dayne Freitag, Matthias Blume, John Byrnes, Edmond Chow, Sadik Kapadia, Richard Rohwer, and Zhiqiang Wang. 2005. New experiments in distributional representations of synonymy. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pages 25–32. Association for Computational Linguistics.

Peter G.W. Glare. 1982. *Oxford latin dictionary*. Oxford univ. press.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018a. Learning Word Vectors for 157 Languages. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hlne Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3843–3847, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).

Edouard Grave, Armand Joulin, and Quentin Berthet. 2018b. Unsupervised Alignment of Embeddings with Wasserstein Procrustes. pages 1880–1890.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany, August. Association for Computational Linguistics.

John Hill. 1804. *The Synonymes in the Latin Language, Alphabetically Arranged; with Critical Dissertations Upon the Force of Its Prepositions, Both in a Simple and Compounded State: By John Hill, LL. D. Professor of Humanity in the University, and Fellow of the Royal Society, of Edinburgh*. James Ballantyne, for Longman and Rees, London.

Bernhard Jussen and Gregor Rohmann. 2015. Historical Semantics in Medieval Studies: New Means and Approaches. *Contributions to the History of Concepts*, 10(2):1–6.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer.

2016. Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT*, pages 260–270.

Thomas K. Landauer and Susan T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211–240.

Susan Leavy, Karen Wade, Gerardine Meaney, and Derek Greene. 2018. Navigating literary text with word embeddings and semantic lexicons. In *Workshop on Computational Methods in the Humanities 2018 (COMHUM 2018), Luasanne, Switzerland, 4-5 June 2018*.

Thang Luong, Richard Socher, and Christopher Manning. 2013. Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Marco Passarotti, Marco Budassi, Eleonora Litta, and Paolo Ruffolo. 2017. The Lemlat 3.0 Package for Morphological Analysis of Latin. In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, number 133, pages 24–31. Linköping University Electronic Press.

Marco Passarotti. 2010. Leaving behind the less-resourced status. The case of Latin through the experience of the Index Thomisticus Treebank. In *7th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages LREC 2010, Valetta, Malta, 23 May 2010 Workshop programme*, pages 27–32.

Marco Passarotti. 2015. What you can do with linguistically annotated data. from the index thomisticus to the index thomisticus treebank. In *Reading Sacred Scripture with Thomas Aquinas: Hermeneutical Tools, Theological Questions and New Perspectives*, pages 3–44.

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 298–307.

Arnošt Skřivan. 1890. *Latinská synonymika pro školu i dum.* V CHRUDIMI.

Rachele Sprugnoli and Sara Tonelli. 2019. Novel event detection and classification for historical texts. *Computational Linguistics*, 45(2):229–265.

Ludwig Von Doederlein and Samuel Harvey Taylor. 1875. *Döderlein's Hand-book of Latin Synonymes*. WF Draper.

Liang-Chih Yu, Jin Wang, K Robert Lai, and Xuejie Zhang. 2017. Refining word embeddings for sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 534–539.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21.