ISSN 2499-4553

IJCoL

Italian Journal of Computational Linguistics

Rivista Italiana di Linguistica Computazionale

Volume 6, Number 1
june 2020

aAccademia university press

Registrazione presso il Tribunale di Trento n. 14/16 del 6 luglio 2016

Rivista Semestrale dell'Associazione Italiana di Linguistica Computazionale (AILC)

Associazione Italiana di Linguistica Computazionale

direttore responsabile
Michele Arnese

isbn PDF 9791280136404

Accademia University Press
via Carlo Alberto 55
I-10123 Torino
info@aAccademia.it
www.aAccademia.it/IJCoL_6_1

IJCoL

Volume 6, Number 1
june 2020

CONTENTS

Building and Comparing Lemma Embeddings for Latin. Classical Latin versus Thomas Aquinas.

Rachele Sprugnoli *
Università Cattolica del Sacro Cuore

Giovanni Moretti*
Università Cattolica del Sacro Cuore

Marco Passarotti*
Università Cattolica del Sacro Cuore

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. 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. The embeddings built upon the two training corpora are compared to each other to support diachronic lexical studies. The words showing the highest usage change between the two corpora are reported and a selection of them is discussed.

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 the study of them, 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 area, 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*[1] and the corpus of Medieval Italian Latinity *ALIM*[2], can now be processed with state-of-the-art computational tools

* CIRCSE Research Centre - Largo Agostino Gemelli 1, 20123 Milano, Italy. E-mail: {rachele.sprugnoli,giovanni.moretti,marco.passarotti}@unicatt.it

1 http://www.perseus.tufts.edu/hopper/

2 http://www.alim.dfll.univr.it/

and methods to provide derivative linguistic resources that enable scholars to exploit the empirical evidence provided by such datasets to the fullest. Such approach is particularly promising given that the quality of many textual resources for Latin, carefully built over decades, is high, just because of to the closed-corpus nature of Latin, which makes every single textual occurrence matter.

Recent years have seen the rise of "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 (Harris 1951; Firth 1961; Lenci 2018).

The range of applications of word embeddings is wide, covering a large span of Natural Language Processing (NLP) tasks, including Word Sense Disambiguation (Iacobacci, Pilehvar, and Navigli 2016), Language Understanding (Li and Jurafsky 2015) and Sentiment Analysis (Socher et al. 2013). While the relevance of word embeddings for the NLP community is crystal clear, their importance for the area of research in the Humanities dealing with ancient languages is not self-evident. Among others, this is due to two main reasons.

On the one side, word embeddings lack for most of the ancient languages. Indeed, building word embeddings for ancient languages is affected by those specific issues that set them apart from modern languages, the most obvious being that word embeddings for modern languages are built upon huge amounts of training data, which are just not available for ancient languages, including Latin. Although Latin is the ancient language provided with the largest textual corpus, the size of such corpus is incomparable with that of most modern languages, often counting several billion words. 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[3] estimated Ancient Greek and Latin production surviving from Antiquity through 600 CE 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 so called Neo-Latin literature, that is, texts dating from the age of Petrarch (1304-1374) to the present day. To get a rough idea of the size of this further Latin data, one should recall that the CAMENA corpus, ie. the most comprehensive collection of Neo-Latin texts, counts about 50 million words.[4]. Other large sources of Latin data are the Latin Wikipedia (*Vicipaedia*)[5], which contains articles on a wide variety of subjects, and the Internet Archive [6], which features a total of 1 billion words in texts published between 200 BCE and 1922 CE (Bamman and Smith 2012).

On the other side, despite their broad spectrum of applications, the use of the (still few) available word embeddings for ancient languages has not entered yet the everyday life of scholars in the Humanities (and, particularly, of classicists), because in most cases they are not put in the condition of interpreting the results that can be

3 https://www.dh.uni-leipzig.de/wo/projects/open-greek-and-latin-project/
4 http://mateo.uni-mannheim.de/camenahtdocs/camena_e.html
5 https://la.wikipedia.org/wiki/
6 https://archive.org/

produced out of such resources. A more strict collaboration is thus needed between computational linguists working on the development and use of word embeddings for ancient languages and scholars in the Humanities, to make the former address to the best the needs of the latter, and the latter really benefit from the results produced by the former.

Lexical analysis is the most obvious area where such collaboration can come true. For instance, Bamman and Burns (2020) show applications of their BERT Latin embeddings on word sense disambiguation and prediction of missing words for textual criticism purposes. More generally speaking, computational linguists can now provide scholars in Classics with bunches of words that share similar contexts of use in large data sets of texts. In turn, classicists are usually asked by computational linguists to evaluate the results of word embeddings. This is, for instance, the case of the recent SemEval-2020 Task 1[7], where 10 annotators with a high-level knowledge of Latin were recruited to judge the sense distributions of a set of target words whose meaning had changed between the pre-Christian and the Christian era according to the literature (Schlechtweg et al. 2020). Beside exploiting the knowledge of classicists to evaluate the information processed by computational linguists (and, for instance, embodied in word embeddings), one step forward consists of going the opposite way. Once evaluated the quality of word embeddings, these can be used as source of information upon which classicists can build new knowledge, thanks to the current possibility of processing and organizing the contextual behavior of the Latin lexicon as used in large corpora. In sight of this, by comparing different approaches to building lemma embeddings from Latin corpora of different era, this paper wants to provide scholars in the Humanities with a set of words that show a different contextual behavior in the diachronic data.

1.1 Our Contribution

We summarize our contribution as follows:

- after describing the related work on Latin word or lemma embeddings and automatic detection of language change (Section 2), we present and evaluate a number of embeddings for Latin built from a manually lemmatized dataset containing texts from the Classical era (Sections 3 and 4);
- we create a new benchmark for the synonym selection task to be used for the intrinsic evaluation of embeddings (Section 5);
- we analyze language change between two corpora collecting texts of two different time periods so to identify those words whose usage has undergone a remarkable change between the Classical era and the Medieval/Christian era, the latter as represented here by a large selection of the works of Thomas Aquinas (Section 6).

This research is performed in the context of the *LiLa: Linking Latin* project (Passarotti et al. 2020), which seeks to build a Knowledge Base of linguistic resources for Latin connected via a common vocabulary of knowledge description following the principles of the Linked Data framework.[8] Our contribution provides the community with new resources to be connected in the LiLa Knowledge Base aimed at supporting data-driven

7 Unsupervised Language Change Detection (https://competitions.codalab.org/competitions/20948).
8 https://lila-erc.eu/

linguistic, literary and socio-cultural studies of the Latin world. The added value of our contribution results from the interdisciplinary blending of state-of-the-art methods in computational linguistics with the long tradition of Latin corpora creation and use: on the one hand we resort to methods and techniques usually applied to modern languages data only, on the other we employ high quality datasets largely used by scholars working on Latin.

2. Related Work

In this Section we summarise two areas of research relevant to our study. First, we present previous work on the development of word or lemma embeddings for Latin (Section 2.1). Secondly, we briefly review recent methodologies proposed for the automatic detection of language change across time periods and between corpora (Section 2.2).

2.1 Word or Lemma Embeddings for Latin

As already mentioned, word embeddings are crucial to many 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, Leskovec, and Jurafsky 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. 2018), 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[9] 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 and Smith (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[10] 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. -**-, *iftudˆ*. The quality of the corpus used to train the Latin word and lemma embeddings available through the SemioGraph interface[11], on the other hand, is high: these embeddings are based on the "Computational Historical Semantics" database, a manually curated

9 For example, the size of the CoNLL embeddings vocabulary is 1,082,365 words.
10 http://www.cs.cmu.edu/~dbamman/latin.html
11 http://semiograph.texttechnologylab.org/

collection of 4,000 Latin texts written between the 2nd and the 15th century AD (Jussen and Rohmann 2015). 18th century Neo-Latin has instead been the focus of the work by Bloem et al. (2020), who generated word embeddings from very small data in the philosophical domain.

Recently, the organization of EvaLatin 2020, the first evaluation campaign totally devoted to the evaluation of NLP tools for Latin, has given a strong impulse to the development of new embeddings to be used for the Part of Speech (PoS) and Lemmatization tasks (Sprugnoli et al. 2020a). Different types of vector representations have been proposed, such as contextualized and treebank embeddings (Straka and Straková 2020), grapheme embeddings (Bacon 2020) and sub-words embeddings (Stoeckel et al. 2020; Celano 2020).

With respect to the works cited above, in this paper we rely on manually lemmatized texts free of OCR errors, we focus on texts of a period not covered by the "Computational Historical Semantics" database (the so-called Classical era) and we test two models to learn lemma representations. It is worth noting that among the previously mentioned studies, only Bloem et al. (2020) have carried out an intrinsic evaluation of the trained Latin embeddings; we, as well, provide both quantitative and qualitative evaluations of our vectors and we release a benchmark built by Latin language experts.

2.2 Language Change Detection

The automatic identification of lexical sense divergence across corpora has become a very popular task in NLP as demonstrated by the organization of dedicated workshops[12] and evaluation exercises at both national and international level.[13] In the literature, different names are proposed for this task: an ever increasing number of works present computational approaches for the detection of semantic shift, meaning change, conceptual change, or usage change, focusing, in turn, on semantic or lexical aspects of language change.

Language change can be detected either across domains or across time, thus pursuing either a synchronic or a diachronic perspective (Schlechtweg et al. 2019). In both cases, the task is usually addressed by applying distributional semantic models, even if other approaches have been explored as well, for example adopting topic-based (Frermann and Lapata 2016) or graph-based models (Mitra et al. 2015).

Approaches using semantic vector spaces mainly differ from each other in terms of semantic representation (e.g., contextualized or non-contextualized embeddings), alignment method (e.g., Orthogonal Procrustes or Vector Initialization) and change detection measure (e.g., Cosine Distance or Local Neighborhood Distance) (Tang 2018; Kutuzov et al. 2018; Tahmasebi, Borin, and Jatowt 2018).

The current state of the art in language change detection has been assessed in the "Unsupervised Lexical Semantic Change Detection" shared task at SemEval 2020 (Schlechtweg et al. 2020). Participating systems were required to identify the meaning

12 See, for example, the two workshops on "Automatic Detection of Language Change" in 2018 (https://languagechange.org/events/2018-sltc-lcworkshop/) and 2020 (https://languagechange.org/events/2020-sltc-lcworkshop/) and the first "International Workshop on Computational Approaches to Historical Language Change" in 2019 (https://languagechange.org/events/2019-acl-lcworkshop/).

13 See task 1 at SemEval 2020 (https://competitions.codalab.org/competitions/20948) and the "Diachronic Lexical Semantics" task at EVALITA 2020 (https://diacr-ita.github.io/DIACR-Ita/).

change over time of a set of target words, taken from the literature, in a multi-lingual setting including Latin, as well as English, German, and Swedish. Results obtained by 33 teams proved that the task was challenging and far from solved, without a single approach valid across languages.

In this paper, following the work by Gonen et al. (2020), we focus on the task of usage change detection in a diachronic perspective. Our aim is to identify words showing a different usage in two corpora providing texts of as many different eras, by using lemma embeddings trained on them. This approach proved to be simpler, more stable and more interpretable than methods requiring vector space alignment.

3. Dataset Description

Our first 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.[14] 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 (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).[15]

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 CBOW 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

14 The composition of the corpus is reported at:
http://web.philo.ulg.ac.be/lasla/textes-latins-traites/.

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

texts and transformed *v* into *u* (so that *vir* 'man' becomes *uir*) to fit the lexicographic conventions of some Latin dictionaries, as for instance Glare (1968), and NLP tools, e.g. LEMLAT (Passarotti et al. 2017). With the minimal number of lemma occurrences set to 5, we obtained a vocabulary size of 11,327 lemmas.

5. Evaluation

Lemma 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.

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

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

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, Dinu, and Kruszewski 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, namely: the target word and another four, one of which is a synonym of the target word and the remaining three are 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 of Latin synonyms (Hill 1804; Dumesnil 1819; Von Doederlein and Taylor

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

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

	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%**

1875; Skřivan 1890), all available online in XML Dictionary eXchange format.[16] 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 the morphological analyzer for Latin LEMLAT v3[17] 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 morphological 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

16 https://github.com/nikita-moor/latin-dictionary
17 https://github.com/CIRCSE/LEMLAT3

Table 4
Examples of the nearest neighbors of rare lemmas.

	***contrudo*/to thrust**	***frugaliter*/thriftily**
FastText-skip	*protrudo**/to thrust forward *extrudo**/to thrust out	*frugalis**/thrifty *frugalitas**/economy
FastText-cbow	*contego**/to cover *contraho*/to collect	*aliter*/differently *negligenter*/neglectfully
Word2Vec-skip	*infodio*/to bury *tabeo*/to melt away	*frugi*/frugal *quaerito*/to seek earnestly
Word2Vec-cbow	*refundo*/to pour back *infodio*/to bury	*lautus*/neat *frugi**/frugal

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 investigation.

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 supposed to be due to the fact that rare words in Word2Vec have few words sharing the same contexts 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 five randomly selected lemmas appearing between 5 and 10 times only in the "Opera Latina" corpus.[18] 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 two of the selected rare lemmas and their neighbors: an asterisk marks neighbors that the 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: 100% of lemmas obtained with that configuration were recognized as similar by the two experts.

18 The lemmas are: *auspicatus* 'augury', *cinnamum* 'cinnamus', *contrudo* 'to thrust', *frugaliter* 'thriftily', *transcribo* 'to copy'.

6. A Diachronic Perspective

Diachronic analyses are particularly relevant for Latin given that its use spans more than two millennia. In order to study language change between the Classical and the Medieval/Christian eras, we performed a computational analysis using two corpora: i.e. "Opera Latina" (see Section 3) and "Opera Maiora", that is the collection of the main works written by Thomas Aquinas in the 13th century. "Opera Maiora" is a set of philosophical and theological treatises 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" corpus. We pre-processed this corpus following the conventions adopted in "Opera Latina": we lower-cased, removed punctuation, and converted *v* and *j* into *u* and *i*, respectively.

In order to detect words that are used differently in "Opera Latina" and "Opera Maiora" and, thus, to enhance corpus-based research, we followed the approach proposed by Gonen et al. (2020) and briefly described below.

6.1 Method

The method of Gonen et al. (2020) (hereafter referred to as NN) is based on the nearest neighbors of words in the embedding spaces trained on two corpora under study, taking into consideration the vocabulary space shared between the embeddings. The usage change of a word is determined by considering the list of its top-*K* neighbors in the embeddings of each of the two corpora: the size of the intersection of these two lists is then computed and words with a smaller intersection are those that potentially have changed more. The output of this procedure is a ranked list of words that starts with the candidates whose usage is most likely to have changed.

The original implementation considers large corpora of several million tokens and a large number of nearest neighbors, setting *K*=1000. In addition, it uses the skip-gram model of Word2Vec with 300 dimensions to build the word representations. Since our corpora, and thus our vocabularies, have a more limited size, we followed the suggestion given by the authors to reduce the *K* and we decided to also test our skip-gram FastText embeddings with 100 dimensions that proved to be the best representation in the evaluations described in Section 5. In order to find the most stable *K* and vector dimensions, we ran the method with different *K*s (100, 200, 500, 1000), different dimensions (100 and 300) and employing both the original implementation, training Word2Vec embeddings, and a new one, taking as input our FastText embeddings. We took the ranks generated by all the runs created by combining different *K*, vector size and vector representation type and calculated the average rank of words common to all ranks. The combination that produced the rank with the lowest standard deviation compared to the average rank was considered the most stable one. The lower oscillation was recorded with *K*=500 and 300 dimensions. Once we found the best settings, we decided to take advantage of both word representations at our disposal (Word2Vec and FastText) by joining the two ranks obtained from each run using those embeddings.

6.2 Results and Discussion

Table 5 shows the top-20 lemmas resulting from the joining method described in Section 6.1; in other words, it reports the words undergoing a substantial usage change in the two corpora in question as detected by both the vector representations used (Word2Vec and FastText).

Table 5
Top-20 detected words. Translation is taken from the Oxford Latin Dictionary (Glare 1968): only the first sense of each lemma, as provided by the dictionary, is reported here.

RANK	COMMON WORDS	RANK	COMMON WORDS
1	*equus*/horse	11	*uirgo*/girl of marriageable age
2	*altus*/tall	12	*sanguis*/blood
3	*sacer*/sacred	13	*lumen*/light
4	*niger*/black	14	*patrius*/belonging to a father
5	*fundo*/to pour (verb IV conj) - to base (verb I conj)	15	*gradus*/step
6	*insto*/to set foot on	16	*purus*/clean
7	*brutus*/brute (adj) - Brutus (proper noun)	17	*nobilis*/familiar
8	*rapio*/to snatch away	18	*membrum*/an organ of the body
9	*multitudo*/abundance	19	*condo*/to insert
10	*tendo*/to stretch	20	*facies*/appearance

In the following Subsections, we focus on a number of words from Table 5, discussing whether and how their usage changes in the corpora, possibly reflecting different meanings. To this end, we refer to the nearest neighbors in the embedding spaces of the two corpora as extracted from the NN method and appearing using both Word2Vec and FastText embeddings.

6.2.1 Equus

The different usage of the word *equus* 'horse' in "Opera Latina" and in Thomas Aquinas reflects both the different topics of the texts collected in the two corpora and the peculiar writing style of the philosopher, who makes use of similitudes to support his reasoning.

In Classical Latin texts, *equus* shows a context of use similar to that of words belonging to the war lexicon, including *hasta* 'spear', *sagitta* 'arrow, shaft', *agmen* 'course, army on the march', *turma* 'troop, squadron (of horses)' and *pugna* 'fight, battle'.

Instead, in the lemma embeddings built upon the collection of the works of Thomas Aquinas, the position of the vector of *equus* reflects its usage in the philosophical reasoning, particularly concerning the comparison and difference between human beings and animals, often in the form of similitude, and with a specific focus on the properties of their intellect and nature. One of the words closest to *equus* is *socrates* 'Socrates', which since Aristotle is often used in philosophy as an example of an individual term and, more specifically, of a rational mortal animal in the Porphyrian tree[19] as opposed to universal terms like *equus*, as an example of irrational mortal animals. It is thus not surprising that among the words showing an usage similar to that of *equus* in Aquinas are *irrationalis* 'irrational', *animalis* 'animate', *indiuidualis* 'individual' (Adj.) and *specificus* 'pertaining to / characterizing / constituting a species'.[20] One example showing a co-occurring usage of the words *equus* and *socrates* in the texts of Thomas Aquinas is the following sentence taken from *Scriptum super Sententiis Magistri Petri Lombardi* (lib. 1 d.

19 The so called Porphyrian tree is a tree-like diagram suggested by the Greek philosopher Porphyry in his "Isagoge" to present the Aristotle's classification of categories through a two-way division of species, which are defined by a *genus*, i.e. an existing definition that serves as a portion of the new definition, and a *differentia*, i.e. the new portion of the definition.

20 It is worth noting that the word *specificus* is provided with a lexical entry in the lexicon of Thomas Aquinas by Deferrari et al. (1948), while it is not reported in Classical Latin dictionaries, like Lewis and Short (1879) and Glare (1968).

21 q. 2 a. 1 co.): "ex hoc enim quod dicitur, Socrates est homo, intelligitur quod non est asinus vel equus" ("from the fact that it is said 'Socrates is a man', it is understood that he is not an donkey or an horse").

6.2.2 Sacer

The usage of *sacer* 'sacred, hallowed' changes greatly in the two collections of texts, following the advent of Christianity as a deep cultural turn that mutates the idea of what is sacred, as well as the sacred rites of a community.

In Classical texts, the usage of the word *sacer* is similar to that of names of gods and goddesses of the Latin pantheon (*ceres* 'Ceres', *iuno* 'Juno', *apollo* 'Apollo'), as well as of places related to sacred rites, like *lucus* 'a sacred grove', *sepulcrum* 'burial place', *nemus* 'wood (consecrated to a deity)', *focus* 'fireplace' but also 'sacrifical altar'.

Instead, the word *sacer* in Thomas Aquinas is closely related to ecclesiastical terms, like *chrisma* 'unction', as well as to Classical Latin words used in the Christian sense, like *ecclesia*, shifting from 'assembly of the people' to 'church', and *scriptura*, changing from 'writing, written matter' to 'Scriptures'. It is worth noticing that also the verb *trado* 'to hand over' (but also, and maybe mainly in Aquinas, 'to transmit') is among the neighbors of *sacer*, showing the important role played by "sacra scriptura" in transmitting and presenting the Christian religion.

6.2.3 Rapio

The wide range of meanings of the verb *rapio* (15 in Glare (1968)) can be reduced to two main ones, namely: the proper meaning 'to take away by force, violently' and the figurative one 'to snatch (with the senses)'. By looking at the nearest neighbors of *rapio* in the embeddings built upon the two corpora in question, we see how one of the two meanings is clearly predominant over the other in each of the corpora.

In the "Opera Latina" corpus, *rapio* shows a usage similar to words like *rapidus* 'flowing violently, swiftly moving', *diripio* 'to tear in pieces', *saeuio* 'to be fierce, to rage', *furo* 'to rage, to be mad' and *cruentus* 'covered with blood'. Instead, in the "Opera Maiora", *rapio* is most similar to *paulus*, the apostle converted when the ascended Jesus appeared to him in a great bright light while he was traveling on the road from Jerusalem to Damascus. Indeed, other words with a usage close to that of *rapio* in the texts of Thomas Aquinas are *exalto* 'to raise, to elevate', *glorifico* 'to glorify' and *eleuo* 'to lift up, to raise'.

The data clearly show that while in the "Opera Latina" corpus, i.e. in Classical Latin texts, *rapio* is mainly used in its proper meaning, in the works of Thomas Aquinas the figurative meaning takes over, also in the light of their theological and philosophical content. Not by chance, the lexicon of Thomas Aquinas by Deferrari et al. (1948) reports for *rapio* the definition "to carry to a state in which the mind is, as it were, freed or raised above the body", which is not provided by Glare (1968), where the definition of the figurative meaning of *rapio* is "to snatch (with the senses)" and "to sweep along (into a state of mind)".

6.2.4 Uirgo

Like *sacer*, also the word *uirgo* 'girl of marriageable age, virgin' has clearly undergone a diachronic usage change from the Classical texts of "Opera Latina" to those of Thomas Aquinas, reflecting the wide impact of Christianity on culture and values in the Western world.

In the Classical texts, the word *uirgo* shows a usage such that it is closely related to words belonging to the semantic field of marriage, like *hymen* (and the derived adjective *hymenaeus*) 'the god of marriage', *nupta* 'wife, bride', *thalamus* 'chamber, marriage-bed, bridal-bed', *maritus* 'nuptial' (Adj.) and 'husband' (Noun). The topic of marriage is strictly bound to that of family, represented by other words with a usage similar to *uirgo* in "Opera Latina", like *soror* 'sister' and *genitor* 'parent'. The proper name of a goddess (*iuno*, 'Juno') is also closely related to the usage of *uirgo* in the corpus.

The situation is remarkably different in the works of Thomas Aquinas, where *uirgo* is found similar to the names (or epithets) of the members of the Holy Family (*maria* 'Mary', *ioseph* 'Joseph', *christus* 'Christ', *saluator* 'saviour'). Like *iuno* was the feminine proper name closest to *uirgo* in Classical texts, *maria* is the one playing the same role in Thomas Aquinas' works. Another semantic field strictly related to *uirgo* is that of nativity, represented by words like *conceptio* 'conception, becoming pregnant', *partus* 'childbirth', *natiuitas* 'nativity' and *uterus* 'womb'.

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. We also described an experiment of automatic language usage change detection based on our embeddings, which resulted in a set of words showing a remarkable usage change across two Latin corpora collecting texts of different time periods.

Our contribution is two-fold. On the one side, our embeddings can be used by the NLP community to automatically process further Latin texts, develop tools (and/or models) for various NLP purposes and build new linguistic resources. On the other, by discussing specific cases of lexical usage change, we show how the application of techniques for developing and comparing embeddings built upon different corpora provides results that can be helpful to those scholars in the Humanities who work in the area of Classical languages and, particularly, deal with issues related to diachronic lexical analysis.

The experiment we presented in the paper concerns the comparison of the lexical usage between a corpus of Classical Latin and the set of the major works of Thomas Aquinas. The data of the latter are taken from the "Index Thomisticus" corpus by father Roberto Busa, one of the pioneers of linguistic computing. One of the last research projects of father Busa (who died in 2011) was the so-called "Lessico Tomistico Biculturale" (Bicultural Thomistic Lexicon) (Busa 1992), aiming at building a new, empirically motivated lexicon of Thomas Aquinas, based on the "Index Thomisticus". Today, resources like the lemma embeddings we built, together with different techniques to compare them with other similar resources based on different data, can support a similar project. Such an organization of the large empirical evidence at their disposal would help lexicographers and philosophers to better manage and benefit from it.

Our embeddings can be re-used to support the development of new linguistic resources. One example is a sentiment lexicon for Latin[21], which was automatically induced from the embeddings trained with the Word2Vec representation, 100 dimensions and the CBOW model, employing a k-NN algorithm implementation (Sprugnoli et al. 2020b). Beside their re-use, several future works are envisaged concerning the embeddings themselves. For example, we plan to develop new benchmarks, like the

21 https://github.com/CIRCSE/Latin_Sentiment_Lexicons

analogy test (Mikolov et al. 2013b) or the rare words dataset (Luong, Socher, and Manning 2013), for the intrinsic quantitative evaluation of Latin embeddings. Also, we will interlink the embeddings with the other linguistic resources for Latin in the LiLa Knowledge Base.[22]

All the resources described in this paper are freely available online: `https://github.com/CIRCSE/Lemma-Embeddings-for-Latin`.

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.[23]

References

Bacon, Geoff. 2020. Data-driven choices in neural part-of-speech tagging for Latin. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 111–113, Marseille, France, May. European Language Resources Association (ELRA).

Bamman, David and Patrick J Burns. 2020. Latin bert: A contextual language model for classical philology. *arXiv preprint arXiv:2009.10053*.

Bamman, David 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.

Baroni, Marco, 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)*, pages 238–247, Baltimore, Maryland, June. Association for Computational Linguistics.

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

Bjerva, Johannes 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, Beijing, China, July. Association for Computational Linguistics.

Bloem, Jelke, Maria Chiara Parisi, Martin Reynaert, Yvette Oortwijn, and Arianna Betti. 2020. Distributional semantics for neo-Latin. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 84–93, Marseille, France, May. European Language Resources Association (ELRA).

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

Busa, Roberto. 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.

22 Embeddings can be published and shared following the Linked Data framework applying the Ontolex Module for Frequency, Attestation and Corpus Information (FrAC) `https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/index.md`.

23 **Authors' Contributions.** This article is the result of the collaboration between the three authors. For the specific concerns of the Italian academic attribution system, Rachele Sprugnoli is responsible for Sections 2, 3, 5 and Marco Passarotti is responsible for Sections 1 and 6.2. Sections 4 and 6.1 were written by Rachele Sprugnoli and Giovanni Moretti while Section 7 is the result of a collaborative writing between the three authors.

Busa, Roberto. 1992. Ermeneutica e traduzione: prospettive di un lessico tomistico 'biculturale'. *Medioevo*, XVIII:3–20.

Celano, Giuseppe. 2020. A gradient boosting-Seq2Seq system for Latin POS tagging and lemmatization. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 119–123, Marseille, France, May. European Language Resources Association (ELRA).

Collobert, Ronan, 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.

Deferrari, Roy J., Ignatius McGuiness, and M. Inviolata Barry. 1948. *A lexicon of St. Thomas Aquinas based on the Summa theologica and selected passages of his other works*. Catholic University of America Press, Washington DC.

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

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

Ehlert, Bret R. 2003. Making accurate lexical semantic similarity judgments using word-context co-occurrence statistics. Master's thesis, University of California, San Diego.

Firth, John Rupert. 1961. *Papers in Linguistics 1934-1951: Repr*. Oxford University Press.

Freitag, Dayne, 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 (CoNLL-2005)*, pages 25–32, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Frermann, Lea and Mirella Lapata. 2016. A bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45.

Glare, Peter GW. 1968. *Oxford Latin Dictionary*. Clarendon Press, Oxford.

Gonen, Hila, Ganesh Jawahar, Djamé Seddah, and Yoav Goldberg. 2020. Simple, interpretable and stable method for detecting words with usage change across corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 538–555, Online, July. Association for Computational Linguistics.

Grave, Edouard, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. 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, Hélène 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).

Hamilton, William L., 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.

Harris, Zellig S. 1951. Methods in structural linguistics.

Hill, John. 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.

Iacobacci, Ignacio, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Embeddings for word sense disambiguation: An evaluation study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 897–907, Berlin, Germany, August. Association for Computational Linguistics.

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

Kutuzov, Andrey, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Lample, Guillaume, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California, June. Association for

Computational Linguistics.
Landauer, Thomas K. 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.
Leavy, Susan, 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)*, Lausanne, Switzerland, 4-5 June.
Lenci, Alessandro. 2018. Distributional models of word meaning. *Annual review of Linguistics*, 4:151–171.
Leonhardt, Jürgen. 2009. *Latein. Geschichte einer Weltsprache*. C.H. Beck.
Lewis, Charlton T. and Charles Short. 1879. *Latin Dictionary [founded on Andrew's Edition of Freund's Latin Dictionary]*. Clarendon Press, Oxford.
Li, Jiwei and Dan Jurafsky. 2015. Do multi-sense embeddings improve natural language understanding? In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1722–1732, Lisbon, Portugal, September. Association for Computational Linguistics.
Luong, Thang, 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, Sofia, Bulgaria, August. Association for Computational Linguistics.
Mikolov, Tomás, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR, Workshop Track Proceedings*, Scottsdale, Arizona, USA, May 2-4.
Mikolov, Tomás, 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, Lake Tahoe, Nevada, Usa, December.
Mitra, Sunny, Ritwik Mitra, Suman Kalyan Maity, Martin Riedl, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2015. An automatic approach to identify word sense changes in text media across timescales. *Natural Language Engineering*, 21(5):773.
Passarotti, Marco. 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, Workshop programme*, pages 27–32, Valletta, Malta, 23 May.
Passarotti, Marco. 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.
Passarotti, Marco, 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*, pages 24–31, Gothenburg, May. Linköping University Electronic Press.
Passarotti, Marco, 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.
Schlechtweg, Dominik, Anna Hätty, Marco Del Tredici, and Sabine Schulte im Walde. 2019. A wind of change: Detecting and evaluating lexical semantic change across times and domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 732–746, Florence, Italy, July. Association for Computational Linguistics.
Schlechtweg, Dominik, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online), December. International Committee for Computational Linguistics.
Schnabel, Tobias, 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, Lisbon, Portugal, September. Association for Computational Linguistics.
Skřivan, Arnošt. 1890. *Latinská synonymika pro školu i dum*. V CHRUDIMI.

Socher, Richard, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October. Association for Computational Linguistics.

Sprugnoli, Rachele, Marco Passarotti, Flavio Massimiliano Cecchini, and Matteo Pellegrini. 2020a. 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).

Sprugnoli, Rachele, Marco Passarotti, Daniela Corbetta, and Andrea Peverelli. 2020b. Odi et Amo. creating, evaluating and extending sentiment lexicons for Latin. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3078–3086, Marseille, France, May. European Language Resources Association.

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

Stoeckel, Manuel, Alexander Henlein, Wahed Hemati, and Alexander Mehler. 2020. Voting for POS tagging of Latin texts: Using the flair of FLAIR to better ensemble classifiers by example of Latin. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 130–135, Marseille, France, May. European Language Resources Association (ELRA).

Straka, Milan and Jana Straková. 2020. UDPipe at EvaLatin 2020: Contextualized embeddings and treebank embeddings. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 124–129, Marseille, France, May. European Language Resources Association (ELRA).

Tahmasebi, Nina, Lars Borin, and Adam Jatowt. 2018. Survey of computational approaches to diachronic conceptual change. *CoRR*, abs/1811.06278.

Tang, Xuri. 2018. A state-of-the-art of semantic change computation. *Natural Language Engineering*, 24(5):649–676.

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

Yu, Liang-Chih, 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, Copenhagen, Denmark, September. Association for Computational Linguistics.

Zeman, Daniel, 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, Brussels, Belgium, October. Association for Computational Linguistics.