Proceedings of 1st Workshop on Language Technologies for Historical and Ancient Languages, pages 84–93
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Distributional Semantics for Neo-Latin

Jelke Bloem, Maria Chiara Parisi, Martin Reynaert, Yvette Oortwijn and Arianna Betti
Institute for Logic, Language and Computation, University of Amsterdam
{j.bloem, m.w.c.reynaert}@uva.nl, {mariachiara.paris, yvette.oortwijn, ariannabetti}@gmail.com

Abstract

We address the problem of creating and evaluating quality Neo-Latin word embeddings for the purpose of philosophical research, adapting the Nonce2Vec tool to learn embeddings from Neo-Latin sentences. This distributional semantic modeling tool can learn from tiny data incrementally, using a larger background corpus for initialization. We conduct two evaluation tasks: definitional learning of Latin Wikipedia terms, and learning consistent embeddings from 18th century Neo-Latin sentences pertaining to the concept of *mathematical method*. Our results show that consistent Neo-Latin word embeddings can be learned from this type of data. While our evaluation results are promising, they do not reveal to what extent the learned models match domain expert knowledge of our Neo-Latin texts. Therefore, we propose an additional evaluation method, grounded in expert-annotated data, that would assess whether learned representations are conceptually sound in relation to the domain of study.

Keywords: distributional semantics, evaluation, small data, philosophy, digital humanities, Neo-Latin

1. Introduction

Christian Wolff (1679-1754)'s philosophical ideas on the so-called 'mathematical method' are deemed greatly influential upon 18th century thinking about science (Frängsmyr, 1975, 654-55). An interesting research question is whether the influence of Wolff's ideas can be more precisely assessed by using a mixed (quantitative, qualitative and computational) approach along the lines of Betti et al. (2019) and Ginammi et al. (2020). In addressing this question, we want to link concepts and terms used to express them using computational techniques, including query expansion based on distributional semantics, information retrieval as a downstream task, and meaning shift analysis built upon this.

The endeavour involves several challenges, starting with (i) building a high-quality, multi-author 18th century philosophy corpus with distinctive characteristics including Neo-Latin texts; and (ii) getting satisfactory distributional semantics models for Neo-Latin. In this paper we report results on (ii), and describe initial steps towards (i). As to (ii), our goal is to evaluate Neo-Latin (word) embeddings learned from tiny data (very small data, i.e. a few sentences, following Herbelot and Baroni (2017)) from the specific domain of philosophy, adapting methods known to work well for this data type, but previously applied to English only (Herbelot and Baroni, 2017; Bloem et al., 2019). We perform two evaluation tasks: 1. compare embeddings learned from a single Vicipaedia definitional sentence to Word2vec (Mikolov et al., 2013) embeddings learned from the full Vicipaedia corpus, and 2. test the consistency of embeddings trained on tiny amounts of topic-specific 18th century Neo-Latin data, initialized using different background corpora.

2. Background

Advances in natural language processing and expanding digital archives have made it possible to analyse old texts in new ways (Hinrichs et al., 2019). Distributional semantics (DS) (Turney and Pantel, 2010; Erk, 2012; Clark, 2015) has emerged as an effective way to computationally represent words and sentences in a way that appears to represent their semantic properties. Along with its prevalence in present-day natural language processing, this aspect makes DS a promising family of techniques for application in text-based fields. The application of DS models to historical languages is however challenging, as large amounts of training data are required (Bengio et al., 2003), while relatively little new digital text is being produced online, in comparison with living languages. Artefacts from digitization processes such as Optical Character Recognition (OCR) may also pose problems. At the same time, philosophers who are interested in Latin texts make accurate studies of concepts and expect high accuracy from the digital tools they use. Application of DS models in this context therefore demands the use of specific methods suited to low-resource languages, small corpus sizes and domain-specific evaluation.

2.1. Latin word embeddings

Latin is a highly inflectional language with words taking many forms depending on features such as case and gender, and language models tend to perform worse on inflectional languages. This effect is greater in n-gram models (Cotterell et al., 2018) due to how each word form is represented separately, leading to a large vocabulary. Word2vec also represents words in this way.

DS models of Latin have only been explored to a limited extent, and never for Neo-Latin texts. In contrast to the more numerous and larger-sized Latin corpora of the so-called *Latinitas Romana*, or Classical Latin (7th cent. B.C.-6th cent. A.D.), Latin corpora of the so-called *Latinitas Nova*, or Neo-Latin (15th cent. A.D.-21st cent. A.D.), also called New Latin when referring specifically to the language, are usually smaller in size,[1] and they often present linguistic variations or new word types in comparison to Classical Latin corpora. For example, the terms *analyticus* (analytic) or *syntheticus* (synthetic) are present only in Neo-Latin, and not in Classical Latin. Various Latin corpora are available. Vicipaedia, the Latin Wikipedia, contains 10.7M tokens of text that has been written in recent years. The Latin Library (16.3M to-

[1]For example, in the LatinISE historical corpus v2.2 (McGillivray and Kilgarriff, 2013), the subcorpus Historical_era_Romana (8,069,158 tokens) is considerably bigger than the Historical_era_Nova one (1,985,968 tokens)

kens) is available in plain text format[2], containing texts from all time periods. There are a few manually annotated treebanks: the Index Thomisticus Treebank (Passarotti, 2019) (354k tokens, 13th century, the works of Thomas Aquinas) based on the Index Thomisticus (Busa, 1974), Perseus (Bamman and Crane, 2011) (53K tokens, Classical Latin) and Latin PROIEL (Haug and Jøhndal, 2008) (Classical Latin and the 4th century Vulgate New Testament translations). These are all partially available in Universal Dependencies format, including tokenization, lemmatization and dependency syntax (Nivre et al., 2016; Cecchini et al., 2018). Furthermore, there is the Late Latin Charter Treebank (Korkiakangas and Passarotti, 2011) (250k tokens, medieval Latin). There is some big data as well, specifically a 1.38 billion token corpus of Latin OCRed text (Bamman and Smith, 2012), a large but rather noisy resource due to mishaps in the OCR and automatic language detection processes.

Some Latin DS models exist: Latin data has been included in large multilingual semantic modeling (Grave et al., 2018) and parsing (Zeman et al., 2018) efforts, using automatic language detection to identify the material as Latin. Another large-scale approach was taken by Bjerva and Praet (2015), who trained embeddings on the aforementioned Bamman corpus (Bamman and Smith, 2012) using Word2vec (Mikolov et al., 2013). Parameters were taken from Baroni et al. (2014), who tuned on an English word similarity resource with models trained on a concatenation of large English-language corpora. The resulting models were not tuned or evaluated for Latin. Manjavacas et al. (2019) applied fastText to the same data to create embeddings for the task of semantic information retrieval, also without tuning, finding that more basic BOW methods outperform it and finding fastText to outperform Word2vec. The only study we are aware of that includes an evaluation of Latin word embeddings is by Sprugnoli et al. (2019), who create lemma embeddings from a manually annotated corpus of Classical Latin, the 1.7M token *Opera Latina* corpus, which includes manually created lemmatization. Sprugnoli et al. (2019) evaluate the lemma embeddings by extracting synonym sets from dictionaries and performing a synonym selection task on them. For a given target term, the cosine distance of its vector to a set of four other terms is computed, one of which is a synonym. To successfully complete the task, the synonym has to be nearer to the target term than the alternative terms. The alternative terms were manually checked to make sure they are not synonyms as well. They find that fastText-based models, which can represent subword units, perform better on this task than Word2vec-based model. They note that this may be due to Latin's heavily inflectional morphology, though when using lemmatized data, the effect of morphology should be limited.

In summary, there are no existing DS models relevant for addressing our research question, as Bjerva and Praet (2015)'s models were not evaluated on Latin and Sprugnoli et al. (2019)'s models were designed for Classical Latin. The relevance of the available corpora for creating Neo-Latin word embeddings is an open question that we will address.

[2] `http://thelatinlibrary.com/`, available as part of the Classical Language Toolkit: `https://github.com/cltk/latin_text_latin_library`

2.2. Tiny data

The application of DS models to Latin involves working with smaller datasets than usual in DS. Some work has been done to evaluate the effect of data size and develop methods suited to learning from less data. Factorized count models have been found to work better on smaller datasets (Sahlgren and Lenci, 2016) compared to the Word2vec family of models (Mikolov et al., 2013). Herbelot and Baroni (2017)'s Nonce2Vec, however, shows that Word2vec can be adapted to learn even from a single sentence, if that sentence is highly informative. In an experiment on a small dataset of philosophical texts (Bloem et al., 2019), this method resulted in more consistent embeddings than a count-based model. The way in which Nonce2Vec can learn from such small amounts of data is by learning incrementally, starting from a semantic *background model* that is trained on a larger corpus, such as all Wikipedia text of a language. Given any term with one or a few sentences of context, that term can be placed into this background model, using nothing but those context sentences as training data. First, a simple additive model (Lazaridou et al., 2017) is used for initialization, taking the sum of the Word2vec background space vectors of all the context words of the target term. This additive model is also used as an evaluation baseline. Next, Nonce2Vec trains the background skipgram model on the context sentences for the target term vector, without modifying the network parameters of the background space[3], with an initial high learning rate, large window size and little subsampling. In this way, Nonce2Vec can learn a vector for a target term based on only one or a few sentences of context, even if that term does not occur in the larger background corpus. As we currently have only tiny amounts of in-domain data, and larger corpora are available that can be used as background (see section 2.1.), we use Nonce2Vec to take distributional information from a general-domain background corpus and further train it on our tiny in-domain dataset.

2.3. Evaluation

Distributional semantic models are typically evaluated by comparing similarities between its word embeddings to a gold standard of word similarity scores based on human ratings, such as the MEN dataset (Bruni et al., 2014) or the SimLex-999 dataset (Hill et al., 2015) for English. However, this is a rarely feasible method in specialised domains and low-resource situations. Not only do such datasets not exist for Latin, but even for English, the meaning of words reflected in these resources may differ from their meaning in the philosophical domain (Bloem et al., 2019).

Evaluation sets can also be created automatically using existing resources. Synonym sets, e.g. from lexical semantic databases, can be used as gold standard data by means of a synonym selection task, which measures how often the nearest neighbour of a vector is its synonym. This method was used for Latin by extracting information from dictionaries (Sprugnoli et al., 2019), but for our use case, this approach

[3] Nonce2Vec can also modify the background model in newer versions (Kabbach et al., 2019), but this can lead to a snowball effect, where the context sentence vectors are significantly moved towards the position of the new context through backpropagation, which would worsen the quality particularly of small models.

may also have the issue of not reflecting domain-specific meanings. General dictionary synonyms may not reflect the way words are used in our target domain. Herbelot and Baroni (2017) evaluate Nonce2Vec by using vectors from a Word2vec model of Wikipedia text as gold vectors. The Word2vec model was in turn evaluated using word similarity scores from the MEN dataset. This evaluation can be conducted for any language in which Wikipedia is available, although for Latin, we do not have a word similarity test collection equivalent to the MEN dataset to evaluate a Word2vec model trained on Vicipaedia.

Some aspects of embedding quality can be measured without a gold standard. The metric of *reliability* quantifies the randomness inherent in some predictive distributional semantic models, and to what extent it can affect the results (Hellrich and Hahn, 2016). Bloem et al. (2019) propose *consistency* as a metric for evaluating low-resource DS models, defining a model as consistent "if its output does not vary when its input should not trigger variation (e.g. because it is sampled from the same text)". The consistency metric computes the ability of a model to learn similar embeddings from different parts of homogeneous data, and does not require 'gold' vectors to compute as it only compares learned vectors to each other. Multiple vectors for a single target term but with different context sentences are trained from identically parametrized models, and compared to each other in terms of nearest neighbour rank and cosine similarity. Higher similarity and nearest neighbour rank between these different vectors of the same target term indicates that the model is more consistent at the level of the domain of text that the context sentences are sampled from (a time period, author, genre, topic etc.). While this measure does not capture all aspects of model quality, it can be used to quantify what model configurations and which background corpora produce consistent embeddings.

To evaluate in-domain term meaning, domain-specific knowledge should be used in the evaluation. Comparative intrinsic evaluation (Schnabel et al., 2015) — i. e. letting users compare and rank terms from a list of nearest neighbours against a query term for semantic similarity — can be used to have experts assess the output of a model, and quantify the outcome. When evaluating models of philosophical concepts, this is not a trivial task. As even domain experts might be unaware of all possible expressions of a concept used by a particular author, constructing ground truths of in-domain key concepts paired off with known terms is necessary for evaluation, as shown by Meyer et al. (2019). This, in turn requires a large in-domain corpus. Although we are currently in the process of constructing a corpus with these exact characteristics, we do not have it yet in a form that is suitable for evaluation based on expert ground truths. If constructed properly in a machine-readable way, such a ground truth would enable automatic evaluation of model output in comparison to the ground truth.

3. Tasks

Considering the constraints on data size and evaluation for our domain, we perform two evaluations of Nonce2Vec on Latin data. The first evaluation aims to replicate Herbelot and Baroni (2017)'s English definitional dataset and evaluation for Latin, and shows us that Nonce2Vec can learn meaning representations from a single sentence that are similar to those learned from a larger corpus. In the second task, we evaluate vectors trained on a tiny dataset composed of sentences from texts relevant to our research question on Wolff's mathematical method. We perform the consistency evaluation of Bloem et al. (2019), while testing different background models for initialization. The second evaluation task shows us that Nonce2Vec can learn word embeddings from these sentences consistently even without access to a background corpus from the target domain.[4]

3.1. Vicipaedia definitional dataset evaluation

We built a dataset of terms and their definitional sentences, following Herbelot and Baroni (2017)'s definitional dataset for English using the same procedure as much as possible. We used Vicipaedia as a source, downloaded and extracted using Witokit[5]. This source was chosen because Herbelot and Baroni (2017) also used Wikipedia and because it is relatively close in time to 18th century Neo-Latin, is large, and is free of OCR errors. The dataset was constructed by taking Vicipaedia page titles containing one word only, taking that page title as a target term and taking the first sentence of the corresponding article as the definitional sentence. The sentences were tokenized using Polyglot[6] and we removed punctuation. We then filtered out target terms that occur fewer than 50 times in Vicipaedia to ensure that they are well-represented in the background model. Herbelot & Baroni used a frequency cutoff of 200 in the UkWaC corpus, but our corpus is smaller so we chose a lower cutoff. We also filtered out terms for which the definitional sentence is shorter than 10 words, to ensure there is some context to learn from. Terms for which the title word does not literally occur in the first Vicipaedia sentence were filtered as well. Occurrences of the target term were replaced by the string '___', ensuring that a new vector is learned for that term. We then randomly sampled 1000 of these terms and sentences, splitting them into 700 tuning and 300 test instances. All of this replicates Herbelot and Baroni (2017)'s extraction procedure for English.

To estimate the quality of the extracted material, we manually checked 99 of the randomly sampled definitional sentences and found that 70 contained proper definitions, 21 contained definitions with additional non-definitional information and 8 did not contain proper definitions. As Herbelot and Baroni (2017) extracted full sentences, definitions with additional information also occur in their sets, so we accept these cases. After updating our automatic extraction procedure, of the 8 non-definitional cases, 3 were excluded by excluding cases with parentheses in the title, 2 were resolved by including words between parentheses in the sentence extraction, 1 is a proper name without definition, and 2 now include a definition but also additional material.

Nonce2Vec can use these definitional sentences to perform one-shot learning of the target term. This newly learnt vector

[4] A branch of Nonce2Vec that includes these evaluations and datasets can be found at `https://github.com/bloemj/nonce2vec/tree/nonce2vec-latin`

[5] `https://github.com/akb89/witokit`

[6] `https://github.com/aboSamoor/polyglot`

can then be compared to the vector produced by a standard (skipgram) Word2vec model trained over the entire Vicipaedia. It is expected that a well-performing system will learn from the definitional sentence a vector that is close to the Vicipaedia vector: their Reciprocal Rank (RR) will be high. We calculate RR between the learned vector and the gold Vicipaedia vector from the background model, over all target terms, and take the resulting Mean Reciprocal Rank (MRR) as a measure of model quality. As a baseline, we use the additive model which just sums context vectors from the background space, following Herbelot and Baroni (2017).[7]

3.2. Neo-Latin dataset evaluation

We built a Neo-Latin dataset consisting of terms and their context sentences. This material is lifted from a small portion (about 20%) of a Neo-Latin corpus that is being used in our ongoing work (Van den Berg et al., ongoing). The full corpus includes 162 books in Latin and 146 books in German published in Germany between 1720 and 1790. We estimate the page count of the Neo-Latin corpus at roughly 40.000. The full corpus has several distinctive characteristics. It is (i) built by a team of experts towards a specific scholarly purpose, that of investigating the concept of *mathematical method* in 18th century Germany; (ii) it presents linguistic variation and vocabulary typical of Neo-Latin corpora (see section 2.1.); additionally, the texts contained in the corpus are more recent in comparison to Neo-Latin corpora from e.g. the 15th century. Another characteristic of our corpus is (iii) that it includes *only* academic philosophy, logic and science in general. In addition to focusing on specific topics and their corresponding technical language, the corpus thus also provides insight into the social context of the authors (Europeans with a deep command of Latin, (writing under) male (names), of a certain age and socio-economic background).

Manual annotations on the Neo-Latin corpus are currently ongoing. They aim at extracting lists of terms expressing certain philosophical concepts relevant to the study of the concept of mathematical method in 18th century Germany, as well as their (functional) synonyms, and the context in which they appear. A selection of contexts get manually typed in full. The Neo-Latin dataset we use in our task is a subset of the full annotation set, and is curated by the same annotator of the full Neo-Latin annotation set, a philosopher by training with knowledge of Latin (Maria Chiara Parisi). The dataset presents – *a fortiori* – the features of the full corpus indicated above and consists of a small, manually-typed and manually-checked set of 30 target terms and, for each term, three sentences (see Table 4) in which the term occurs. The target term (column 1) is replaced in the snippets (column 2, 3 and 4) with '___'. The Neo-Latin *corpusculum* we use is a tiny, but sufficient set of data to test the consistency of Neo-Latin word embeddings.

As we do not yet have the full corpus in a suitable machine-readable format, we cannot perform the same evaluation as for the definitional dataset, but we can measure vector *consistency* (Bloem et al. (2019), see 2.3.). We can use an out-of-domain background corpus, such as Vicipaedia, for initialization, in order to use Nonce2Vec to model these terms. Note that, doing this, we can no longer evaluate the resulting vectors by comparing the learned vectors to those from the background corpus. The background corpus is text of a different domain than 18th century mathematical text, and may not even contain the core terms from these works, or it may use them in a different way. Thus, unlike in Herbelot and Baroni (2017)'s Wiki definitions evaluation setup, vectors based on an out-of-domain background corpus cannot serve as a gold standard for vectors from our domain.

The consistency metric (Bloem et al., 2019) evaluates the stability of vector spaces generated by a particular model on a homogeneous dataset extracted from a particular domain of text, without a gold standard. In our case, the model is Nonce2Vec, and the homogeneous dataset is our tiny Neo-Latin mathematical method subset. Consistency is computed by measuring the similarity between vectors of the same word, trained over different samples of text (the sentences from the dataset). We can use this metric to compare different configurations of Nonce2Vec on the task and see which one results in more consistent embeddings. In particular, we are interested in trying different background models for initializing the Nonce2Vec vectors, trained on different background corpora. We hypothesize that a background model that leads to higher consistency scores on this task with our Neo-Latin dataset provides a better initialization for our in-domain term vectors. Such a model, we might conjecture, contains more of the relevant vocabulary, used in a more similar way to that of our texts.

4. Results

4.1. Definitional evaluation

In the first evaluation, we compare vectors trained on Vicipaedia definitional sentences to vectors from the Vicipaedia background model, for the same target term. We first train a standard Word2vec model on Vicipaedia, which Nonce2Vec does using the Gensim (Řehůřek and Sojka, 2010) implementation of Word2vec. While Herbelot and Baroni (2017) do not tune this model, as Vicipaedia is smaller than the English Wikipedia they use, we try to change the default parameters to accommodate this. We find that a higher learning rate ($\alpha = .01$), increased window size (15) and higher subsampling rate (1^{-4}) provides better results on our tuning set. Next, we tune and run Nonce2Vec on our Latin definitional dataset, using the background model for initialization and as the sum baseline. We performed a grid search of the same parameter space as Herbelot and Baroni (2017) do, containing different learning rates ($[0.5, 0.8, \mathbf{1}, 2, 5, 10, 20]$), the number of negative samples ($[\mathbf{3}, 5, 10]$), the subsampling rate ($[500, 1000, \mathbf{10000}]$), and window size ($[5, 10, \mathbf{15}, 20]$). The subsampling rate decay ($[1.0, 1.3, \mathbf{1.9}, 2.5]$) and window decay ($[1, 3, \mathbf{5}]$) are not relevant when training vectors on single sentences. Bold values are the best performing values in Herbelot and Baroni (2017).

Using the tuned Vicipaedia background model and applying it to the test set, the best performance is obtained for a window size of 5, a learning rate of 0.5, a subsampling rate of 500, and 3 negative samples. The lambda parameter was set

[7] We run the Nonce2Vec algorithm without the notion of informativeness incorporated by Kabbach et al. (2019), as that option requires the use of an additional language model.

Model	MRR	Median rank
N2V-best	0.01936	251
N2V-defaultbg	0.15832	1866
N2V-default	0.00410	5736
Sum	0.01263	322

Table 1: Results on definitional dataset

to the default 70. Table 1 shows results using these tuned parameters (N2V-best) and the default Nonce2Vec parameters from the English experiment (N2V-default) as compared to the sum baseline (Sum). The N2V-defaultbg result uses our tuned N2V parameters, but with the default background model parameters, and the N2V-default result uses default parameters from Herbelot and Baroni (2017) both for the background model and for training on the definitional data. On the test instances, we find that N2V shows an improvement over the simple additive model baseline. As shown in Table 1, the median rank of the gold vectors for our test instances is 251, out of 14,049 neighbours (the vocabulary size). For English, Herbelot and Baroni (2017) report a median rank of 623. While this number appears worse than our score, this metric is sensitive to vocabulary size: their English model has a vocabulary of 259,376 types due to the larger corpus, and ranking high is more difficult when there are more other vectors to rank. The Mean Reciprocal Rank (MRR) measure is 0.019 on the Latin definitions but 0.049 on the English definitions, showing that the nearest neighbours of the gold Wiki vectors rank higher among the nearest neighbours of the learned definitional vector for English than for Latin.

4.2. Neo-Latin consistency evaluation

Recall that for the Neo-Latin data that pertains to our philosophical research question, we do not have gold vectors, as there is no background corpus for our domain yet. Instead, we compute consistency between vectors trained over different context sentences of the same target term (shown in Table 4). We experiment with initializing our vectors based on models trained from various background corpora with various model parameters, in order to find out what background model leads to more consistent results for our domain of Latin text. As background corpora, we use the Vicipaedia, Latin Text Library, Latin Treebanks and Bamman corpora described in section 2.1. The Latin Text Library corpus was tokenized using Polyglot in the same way as the Vicipaedia corpus. The Bamman corpus was tokenized and lowercased by Ucto (van Gompel et al., 2017). Punctuation was removed and, as these may be disruptive to distributional models, we let Ucto replace items that are less lexical, such as numbers of any type, dates, etc. by class labels. Of the treebanks, we use the Universal Dependencies versions of the Index Thomistius Treebank (165K tokens), the Perseus LDT (29K) and Proiel (200K).

For each background model, we compute consistency metrics over the vectors learned by Nonce2Vec of all 30 Neo-Latin target terms. We have three vectors per term, one from each context sentence, and compute the metrics between all pairs of the three vectors ($\vec{a_1}$-$\vec{a_2}$, $\vec{a_2}$-$\vec{a_3}$, $\vec{a_1}$-$\vec{a_3}$). This evaluation data is shown in Table 4. We consider two metrics for comparing a pair of vectors $\vec{a_1}$ and $\vec{a_2}$: by similarity, where a higher cosine similarity indicates more consistency, or by nearest neighbor rank, where a higher rank of $\vec{a_1}$ among the nearest neighbors of $\vec{a_2}$ indicates more consistency. Every vector in the background model, as well as $\vec{a_2}$, is ranked by cosine similarity to $\vec{a_1}$ to compute this rank value.

Model	cos-sim	rank	vocab
bamman-c50-d400	0.701	**47.5**	901K
bamman-c50-d100	**0.776**	202	901K
lattextlib-c50-d400	0.332	604	24.7K
lattextlib-c50-d100	0.450	1279	24.7K
lattextlib-c20-d400	0.505	**75**	50.4K
lattextlib-c20-d100	**0.621**	301	50.4K
vicipaedia-c50-d400	0.482	103	14.0K
vicipaedia-c50-d100	0.603	219	14.0K
vicipaedia-c20-d400	0.551	**47.7**	30.4K
vicipaedia-c20-d100	**0.674**	244	30.4K
treebanks-c50-d400	0.133	292	810
treebanks-c50-d100	0.165	286	810
treebanks-c5-d400	0.298	1103	7.3K
treebanks-c5-d100	**0.390**	703	7.3K

Table 2: Consistency metrics on our Neo-Latin dataset using Nonce2Vec, initialized with various background models.

We use the same Nonce2Vec parameters across all experiments: the ones that performed best in our definitional evaluation (section 4.1.). We experiment with background models with different dimensionality: d400 (the Nonce2Vec default) and d100 (found to perform better by Sprugnoli et al. (2019) on lemmatized Latin data). We also vary the frequency cutoff, as when working with smaller data, we may wish to include more words even if they are infrequent. We try a cutoff of 50 (c50), the nonce2vec default, and c20 or c5 depending on the size of the corpus. The results of Nonce2Vec with the different background models are listed in Table 2. We observe that the most consistent vectors are obtained using the largest dataset as a background corpus, the Bamman corpus. Using the largest Bamman model (bamman-c50-d400), we find that different vectors for the same term trained on a different sentence are on average rank 47 in each other's nearest neighbours, out of a vocabulary of 901K types, computed over all 30 test instances. On average, the cosine similarity between these vectors is 0.7. Among the 90 total comparisons between the 3 vectors for the 30 target terms, there were 59 cases where both target term vectors were each other's nearest neighbour (65.6%), with a greater cosine similarity to each other than to any of the other 901K words in the vocabulary. This is an impressive score with a vocabulary of almost a million words. The best-performing Wiki model, with a lower frequency cutoff (vicipaedia-c20-d400) achieves a similar average rank among a vocabulary of 30.4K types, and 51% of comparisons have rank 1 consistency. The cosine similarities are lower, though (0.55). On their synonym detection task for Classical Latin, Sprugnoli et al. (2019) achieve an accuracy of 86.9%, but here, the model only needs to choose between four alternative words, instead of almost 1 million. Furthermore, we observe

Term		$\vec{a_1}$ NNs	$\vec{a_2}$ NNs
genus	1	essentialis	demonstrabilia
	2	metaphysica	quidditative
	3	substantialitas	universaha
conceptus	1	expucetur	possibiles
	2	demonstrabilia	universaliores
	3	universaha	aliquee

Table 3: Qualitative examination of some nearest neighbours of target term vectors computed over two different context sentences of those terms.

that for the bamman-c50-d400 model, the average rank of a target term vector from the background model among the nearest neighbours of the learned Neo-Latin vector for that same term is 50,737 with a cosine similarity of 0.41. This shows that the model does learn from the Neo-Latin data, deviating from the background vector, and does not achieve consistency simply by learning nothing consistently.

Generally, we see in Table 2 that a lower word frequency cutoff (keeping a larger vocabulary) leads to more consistent results. All of this indicates that more background data leads to more consistent vectors on our Neo-Latin data. The Vicipaedia-based models slightly outperform the Latin Text Library-based models, despite their smaller vocabulary. This shows that data size is not the only factor — similarity to our target domain may also be relevant here, as Vicipaedia data may be closer to Neo-Latin scientific text than the contents of the Latin Text Library. Lastly, the models based on the small Classical Latin treebanks perform worst, a corpus that is not only small but also highly varied.

These results show that the Bamman models lead to more consistent embeddings on our data, even though they are based on rather noisy data. We have a closer look at this result by cherry-picking some examples. Table 3 shows the three nearest neighbours for two vectors each for the target terms *genus* (kind) and *conceptus* (concept). $\vec{a_1}$ is trained over the first context sentence for this term from our dataset, and $\vec{a_2}$ over the second. For *genus*, most of these look reasonable — certainly, *essentialis* (essential), *quidditative* (relating to the essence of someone or something) and *substantialitas* (the quality of being substantial or having substance) are semantically related to *genus* in the context of the mathematical method. *Universaha*, while related, is an OCR error (*universalia* (universals)). In this case, the two vectors are also each other's nearest neighbours, so the results for this term are consistent. The nearest neighbours of *conceptus*, on the other hand, are not a very good result. To start, the additive model initialization from the background model for *conceptus* $\vec{a_1}$ has as its nearest neighbours the words *sdbygoogle*, *ìbygoogic* and *digfeedbygoogle*, clearly Google Books artifacts. After training, the nearest neighbours are as listed in Table 3: they have improved compared to the initial additive vector's neighbours and are now vaguely ontopic, but still full of OCR errors. This shows that consistent results are not necessarily of high quality in other respects.

5. Discussion

Our definitional dataset evaluation has shown that Nonce2Vec can learn Latin word embeddings from a single definitional sentence, though slightly less well than it can for English. This is likely because the task of training a DS model is harder on Latin text due to the highly inflectional nature of the language and the smaller size of the Latin Wikipedia. There is less statistical evidence for the usage patterns of more different word forms.

Our Neo-Latin evaluation has shown that Nonce2Vec can consistently learn Neo-Latin word embeddings for terms relevant to a certain concept (i.e. the mathematical method), without access to a background corpus from this domain and without tuning on the consistency metric or Neo-Latin data. The evaluation demonstrates that this method can be used even when nothing but a limited number of sentences is available for the target domain. This is likely due to transfer of word usage information from the general-domain background corpus to the domain-specific sentence context, caused by the way in which Nonce2Vec initializes vectors based on a background corpus. At least two factors may affect the outcome: the size of the background corpus, and how similar it is to Neo-Latin text. Since lack of high-quality corpora in the relevant domain and lack of expert ground truths are typical features of research in low-resource settings, the relevance of our result becomes clear. It is useful in such settings to know that Nonce2Vec learns even from very tiny Neo-Latin corpora – *corpuscula* –, as long as background corpora are available, and that the latter can even be (a) in a different variety of the same language (b) noisy, as long as they are large. Based on this finding, tools that allow information retrieval and visualization using DS models (e.g. BolVis, van Wierst et al. (2018)) can be developed for Latin and applied to digital versions of the relevant texts, in order to find passages relevant to particular research questions in the history of ideas (Ginammi et al., 2020).

Clearly, however, to the aim of addressing our research question on the mathematical method with appropriate scholarly precision, high-quality Neo-Latin word embeddings based on data that is relevant to our concept of interest will be necessary. We encountered several issues related to the morphology of Latin. Among the target terms automatically extracted from Wikipedia, there were many proper names, as they are less affected by morphology. They occur more frequently in their lemma form and are more likely to pass frequency cutoffs. Other Wikipedia lemmas are not frequently used in their lemma form in natural text. In our Neo-Latin dataset, multiple sentences containing the same word form are scarce for the same reason — important terms can be inflected in many ways and each form will get a distinct vector in a standard Word2vec model. Lemmatization has been shown to improve language model performance on highly inflected languages. (Cotterell et al., 2018).

For this reason, Sprugnoli et al. (2019) used lemma embeddings instead of word embeddings. They were able to do this by having a manually lemmatized corpus. For Nonce2Vec, to create lemma embeddings, any background corpus used would have to be lemmatized. Of the corpora we used, only the small treebank corpora that mostly contain Classical Latin contained lemmatization, and none of the better-performing larger corpora exist in lemmatized form. While lemmatizers exist (see Eger et al. (2015) for an overview and evaluation on medieval church Latin) evaluation is costly

and results may vary across different varieties of Latin. Still, for our type of research questions lemmatization carries natural benefits, because, as philosophers focussing on meaning change and concept drift, we are interested in studying concepts independently of the morphological variants of the terms expressing them. In future work, the issue could be addressed with an extrinsic evaluation on our tasks and evaluation across Latin varieties in the context of the EvaLatin shared task (Sprugnoli and Passarotti, 2020).

Despite impressive consistency scores, we also saw that other aspects of the quality of these embeddings may be lacking. Using the top-scoring Bamman model for initialization, we observe many OCR errors among the nearest neighbours of our learned Neo-Latin vectors. This is cause for concern, as Word2vec models based on this same data have already been used in a study of concepts in the works of Cassiodorus (Bjerva and Praet, 2015). We must therefore consider in what ways our evaluation is incomplete. The consistency evaluation does not capture all aspects of embedding quality: after all, a model can be consistently bad as well as consistently good. The definitional evaluation we conducted is only grounded in a larger Word2vec model (the background model) which has not been evaluated for Latin. We also cannot just assume that this model works well on Latin just because it works well on English — as illustrated by the fact that in most of our experiments, the English parameter settings did not perform well on the Latin data. This uncertainty leads us to propose an additional evaluation that is directly grounded in domain expert knowledge, to test whether the learned Neo-Latin word embeddings are not only consistent, but also conceptually sound.

5.1. Grounding the evaluation

To identify whether the word embeddings are consistently good or consistently bad, we need to evaluate them by comparing the domain expert's knowledge of the philosophical data with the embeddings. In Meyer et al. (2019), we propose a first step towards this form of evaluation for a 20th century English corpus of the philosophical works of Quine. For this corpus, we semi-formally defined the relations of some key terms to other terms (e.g., in Quine's oeuvre, *denotation* signifies a relation between a *general term* in language and (*physical*) *objects* in the ontology). By defining these interrelations between terms in the corpus, the expert knowledge of the meaning of a term within the corpus is reflected by how the term relates to other terms. In the case of our Neo-Latin corpus, the domain expert identified that *definitio* (definition) and *axioma* (axiom) are functional synonyms of *principium* (principle). Similar to the task discussed above, to successfully complete this task, the cosine distance of the vector of a given target term has to be nearer to the vectors of their functional synonyms than alternative terms. In the case of *principium*, *definitio* and *axioma*, the cosine distance of the vectors of these terms are expected to be nearer to each other than to other terms. Such a conceptual evaluation grounded in expert knowledge provides a method to evaluate word embeddings intrinsically and, thereby, the quality of their consistency.

5.2. Conclusion

Our results show that consistent Neo-Latin word embeddings can be learned by using methods that are designed to handle tiny data. These methods have not been applied to Latin before. Nonce2Vec might be a good DS model to use in such low-resource settings, although further evaluation and refinement is necessary, in particular in the context of humanities research. In addition, we demonstrate and discuss evaluation methods appropriate for our task. Using both a grounded evaluation and a consistency evaluation can tell us to what extent the learned vectors represent the conceptual distinctions we are interested in, and to what extent they can be learned consistently from the same text source. We have great plans for the future. We are actively digitizing a comparable German and Neo-Latin corpus of philosophical works. We seek to cooperate with existing initiatives and intend to add value to available collections. For e.g. the Bamman corpus this will entail improving the overall text quality by applying fully automatic OCR post-correction as provided by Text-Induced Corpus Clean-up or TICCL (Reynaert, 2010). To equip TICCL for appropriately handling Latin, we will apply the TICCLAT method (Reynaert et al., 2019) for linking morphologically related word forms to each other, to their diachronic and their known typographical variants. This follows from our observation that there is much room for improvements in embedding quality by having lemmatized and cleaned datasets and background corpora. Tiny data methods can also be further explored, as recent work incorporating a notion of informativity and more incrementality into Nonce2Vec (Kabbach et al., 2019) and recent context-based approaches outperforming Nonce2Vec on the English definitional dataset (e.g. Schick and Schütze (2019)) was not explored here. Having high-quality embeddings learned from historical text, and downstream applications that make use of them, will help us in obtaining large-scale evidence for research questions in the history of ideas that is impossible to obtain otherwise.

6. Acknowledgements

We thank the UvA e-Ideas group for their fruitful discussion of a draft of this paper. We also thank the anonymous reviewers for their time and valuable comments. This research was supported by Arianna Betti's VICI grant *e-Ideas* (277-20-007) financed by the Dutch Research Council (NWO) and by Human(e)AI UvA grant *Small data, big challenges*. Yvette Oortwijn has also been supported by *CatVis* (NWO research grant 314-99-117). M. Reynaert is also affiliated to the Tilburg School of Humanities and Digital Sciences, Tilburg University.

Appendix: Neo-Latin evaluation dataset

We here present the Neo-Latin evaluation dataset, non-preprocessed for legibility. Best scores are shown. Provenances of the snippets are documented in the metadata to the online distribution of the experimental data.[8] Shown are smaller excerpts of the longer snippets in the actual dataset.

[8] https://github.com/bloemj/nonce2vec/tree/nonce2vec-latin

Target	PoS	Snippet 1	Snippet 2	Snippet 3	AR	AD
Mathematica	A	methodum meditationis, et inveniendi, et probandi, potest huc commode referri methodus ___	Unde reliquis omnibus praeferenda methodus ___	Ubi methodus ___ adhibetur, ibi & principia sunt indubitata, & modus concludendi legitimus	1	0.848
Mathesin	N	Hoc insuper ipsum me circa ___ particulatim reverä ?	Quamobrem ut ___ huic instituto accommodationem efficierem, omnes disciplinas in eum ordinem digessi.	medium as scientiam perveniendi certissimum est accurata & indefessa matheseos tractatio & methodi ibidem observatae ad alia extra ___ obvia applicatio	1	0.793
Wolfius	PN	___ certe totum opus replevit meris axiomatibus, & postulatis, & observationibus, de quibus nemo facile dubitaverit	Accuratius ergo Illustris ___ in brevi comment. de meth. math. Elementis math. praemissa 30 afferit.	Quicunque itaque Philosophus hoc intelligit, ___ suum sicuti minime concedet.	1.6	0.678
Methodo	N	Possunt etiam quaedam themata tractari ___ mathematica, illa scilicet, quae clara & evidentia principia admittunt.	Methodus demonstrandi synthetica eadem quoque, cum ___ ratiocinandi, est	___ mathematica itaque ad scientiam & subiective & obiective talem possumus pervenire.	1	0.738
Universalis	A	Methodus ___ dirigit totam disciplinam, ejusque disponit.	ac propterea haud immerito elementa matheseos ___ dici queant, suo loco prolixius docebitur.	Methodus ___ dirigit totam disciplinam, ejusque disponit.	115	0.680
Disciplinae	N	quia ex eo certae ___ conclusiones omnes sunt probandae.	qumm pleraeque ___ aliae, quae illam quidem admittant, res solo intellectu comprehensibiles tradant.	Mathesi enim accurate loquendo non est una disciplina, sed ex variis ___ derceptae particulae quantitatem subiecti in unaquaque tractantes	1	0.750
Axiomata	N	Inveniuntur ___ duobus fontibus adhibitis, vel per inductionem, vel per definitiones	Quae enim immediate ex definitiones fluunt, propositiones, si theroreticae sunt, ___, si practicae, POSTULATA, vocantur.	secundò has ipsas definitiones in se considerabo, & hinc deductas proprietas appellabo ___.	20	0.721
Postulata	N	___ ,seu principia generalia, non quidem evidentia ac per se nota, sed tamen aliunde certa aut probata.	Propositiones, quae ex definitione unica adeoque citra ratiocinorum ambagem, manifestae sunt, dicuntur axiomata, si theoreticae fuerint; ___, si praticae.	tum ponantur axiomata & ___, & tandem ipsae demonstrationes subiugantur.	1	0.779
Sequitur	V	cuicunque competit definitum, ei competat definitio, et quicquid ex definitione ___, vel alias de definitio praedicatur.	Sponte sua hinc ___, conclusiones ex hisce principiis derivari posse, quarum inde probetur veritas.	Propositio, quae ex unica definitione ___, legitime tornata statim tanquam vera terminis intellectis patet, modo illa definitio adducatur	4.3	0.696
Definitio	N	Ad cognitionem tria requiruntur: ___, DIVISIO, DEMONSTRATIO	Est autem ___ nihil aliud, quam rei naturae, ex qua ipsa componitur, commoda explicatio.	Et ___ quidem est propositio rem ita determinans	129	0.612
Definitum	A	Notas ad ___ agnoscendum & ab aliis discernendum sufficientes praebent genus atque differentia specifica.	Id quod definitur, vocatur ___	definitionem & ___ esse terminos diversos non solum quoad voces significantes sed etiam quoad conceptus significatos & objectivos	8.6	0.639
Notio	N	___ completa est quae obiecto, omnes ideas in quas necessario resolvitur substituit.	___, quam habemus, aut solas rei notas continet	definitio dicitur, quae est ___ distincta completa speciei vel generis.	6.7	0.652
Ideas	N	Unde simul patet, ___ simplices, eo quod ex aliis non componantur, non quoque posse definiri.	___ mediatas simplices nominabo, quibus opponuntur ideae mediatae compositae	omnes ___ claras quodammodo simplices, distinctas vero semper compositas simul, esse.	1	0.746
Conceptus	N	Definitio, generalissima accepta, est ___ definiti naturam explicans	primò omnes possibiles primo ___, ex quibus formantur reliqui, redigam in ordinem, atque imposterum Definitiones nominabo	quod denique continuum sic, unde ___ completus hic est quod sit extensum ab uno loco ad alterum.	1.7	0.673
Rationem	N	quod vero ___ sui in altero habet principiatum vocatur.	___ unius rei ex altera observantur	haec differentia inter ___ sufficientem & determinantem est ratio	1	0.743
Cognitionem	N	Ad ___ tria requiruntur: DEFINITIO, DIVISIO, DEMONSTRATIO	DE RERUM ___. III. Ad cognitionem tria requiruntur: DEFINITIO, DIVISIO, DEMONSTRATIO.	intellectum ad veram rei ___ ducere debet.	1000	0.652
Simplices	A	Deinde ideae vel ___ sunt, in quibus nihil mente dividere possumus	Res ___, quae nullas habent partes; definiri non possent.	Unde simul patet, ideas ___, eo quod ex aliis non componantur, non quoque posse definiri.	1	0.736
Compositae	A	secundus gradus continet conditiones universales demonstrandi, quae sunt conditiones ___ ex definitionibus principiorum.	Etenim si omnes ideae, quae essentiam ideae ___ constituunt, evoluantur, definitio non potest non esse adequata definitio.	ideas mediatas simplices nominabo, quibus opponuntur ideae mediatae ___,	1	0.757
Genus	N	Definitio Essentialis Metaphysica seu Logica est quae datur per ___ & differentiam.	ideas mediatas compositae, propositio qua per ___ & differentiam specificam declarantur	DEFINITIO, est Oratio explicans natura rei.1. PRIMARIA, quae affert ___ & differentiam	2.3	0.729
Differentia	N	ideam, & quae ad essentiam rei primario pertinet, qua a reliquis omnibus distinguitur, comprehendere debeat, adeoque constare ex genere & ___ specifica.	Partes autem rei Metaphysica sunt Genus, & ___.	Sed attributum substantive expressum dicitur genus, adiective expressum ___.	90	0.608
Principia	N	certa & evidentia axiomata, quae non probantur, sed supponuntur a scientiis, ideoque vocantur prima ____.	Et hoc ipso etiam efficit, ut ex ea ___ clara & evidentia derivari queant.	Quoniam itaque ___ ista immota sunt definitiones, axiomata & experientiae clarae.	1.3	0.726
Terminos	N	definitionem & definitum esse ___ diversos non solum quoad voces significantes sed etiam quoad conceptus significatos & objectivos	Scientia, cujus vult principia invenire, & aliqua notitia habita illius ponit aliquos ___ principiorum	quod sit indemonstrabile, ut possit a quodlibet, ___ modo intelligente, ob evidentiam cognosci.	1	0.729
Intellectus	N	definitio, divisio, argumentatio sunt recta ___ operatio.	definitio, divisio, argumentatio sunt recta cognitio ___.	nexu necessario et evidenti, demonstrantur, ut de illarum veritatem ___ nullo modo dubitare possit.	6.3	0.765
Veritas	N	per axiomata relatio inter certas ideas ostenditur, ut nova inde cognoscatur ___.	a quorum evidentia __ totius discursus continua serie deducitur	Quomodo vocatur illa ___, a qua alia procedit vel dependet?	3	0.611
Propositiones	N	Et huinc Axiomata vocantur etiam Maximae, subintellige, ___	quibus verae ___ possint etiam perfecta fieri & accuratae, quibus nihil, quo fiant meliores, addii posset	Principia dirigentia, sunt axiomata, id est ___ per se notae	1	0.699
Scientia	N	Logica est ___ veritatis	___ est cognitio certa & evidens rei per causam &c.	Primo dicitur ___, quia conclusiones certas ex principiis certis deducit.	2.6	0.659
Theoremata	N	Propositionem vero theoreticam ex pluribus definitionibus inter se collatis erutam, ___ appellant.	Si plures definitiones inter se contenderis: facile reperies ___ & PROBLEMATA.	cum videlicet principia sunt vocantur vel axiomata vel ___	6	0.593
Demonstrationes	N	Principia ___ sunt definitiones	tertius gradus continet ___ universales ex principiis & conditiones universalibus productas	___ autem legitima consequentia ex iis tandem deducant propositiones per se non manifestas.	12.6	0.617
Derivari	V	universalissimum principium est constituendum, a quo cetera omnia facili negotio possunt ___.	Qua ratione autem axiomata ex definitione ___ queant ostendit.	ac quaevis alia requirit methodus, plura possint hinc ___, quam per aliam quamcumque methodum	1	0.725
Demonstrari	V	quod aliunde ___ potest, ac debet, cuius id eo duplex est respectus, alius ad conclusionem, alius ad principium prius clarius	conclusiones ex hisce principiis ___ posse, quarum inde probetur veritatis.	nulla est alia propositio per quam ___ possit veritas hujus propositionis tandem est demonstrativa caeterarum.	1	0.756

Table 4: Neo-Latin consistency evaluation set. Legend: AR = Average Rank, AD = Average Distance, Part-of-speech tags: A = adjective, N = noun, PN = proper noun, V = Verb.

7. Bibliographical References

Bamman, D. and Crane, G. (2011). The ancient Greek and Latin dependency treebanks. In *Language technology for cultural heritage*, pages 79–98. Springer.

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

Baroni, M., Dinu, G., and Kruszewski, G. (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.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Betti, A., van den Berg, H., Oortwijn, Y., and Treijtel, C. (2019). History of Philosophy in Ones and Zeros. In Mark Curtis et al., editors, *Methodological Advances in Experimental Philosophy*, Advances in Experimental Philosophy, pages 295–332. Bloomsbury.

Bjerva, J. and Praet, R. (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.

Bloem, J., Fokkens, A., and Herbelot, A. (2019). Evaluating the consistency of word embeddings from small data. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 132–141, Varna, Bulgaria, September. INCOMA Ltd.

Bruni, E., Tran, N.-K., and Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research (JAIR)*, 49(1-47).

Busa, R. (1974). *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*.

Cecchini, F. M., Passarotti, M., Marongiu, P., and Zeman, D. (2018). Challenges in converting the Index Thomisticus Treebank into Universal Dependencies. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 27–36.

Clark, S. (2015). Vector space models of lexical meaning. *The Handbook of Contemporary semantic theory*, pages 493–522.

Cotterell, R., Mielke, S. J., Eisner, J., and Roark, B. (2018). Are all languages equally hard to language-model? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 536–541, New Orleans, Louisiana, June. Association for Computational Linguistics.

Eger, S., vor der Brück, T., and Mehler, A. (2015). Lexicon-assisted tagging and lemmatization in Latin: A comparison of six taggers and two lemmatization methods. In *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 105–113, Beijing, China, July. Association for Computational Linguistics.

Erk, K. (2012). Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10):635–653.

Frängsmyr, T. (1975). Christian Wolff's Mathematical Method and its Impact on the Eighteenth Century. *Journal of the History of Ideas*, 36(4):653–668.

Ginammi, A., Bloem, J., Koopman, R., Wang, S., and Betti, A. (2020). Bolzano, Kant and the Traditional Theory of Concepts - A Computational Investigation [R&R 16 dec 2019 for volume under contract]. In Andreas de Block et al., editors, *The Dynamics of Science: Computational Frontiers in History and Philosophy of Science*. Pittsburgh University Press.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. *arXiv preprint arXiv:1802.06893*.

Haug, D. T. and Jøhndal, M. (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.

Hellrich, J. and Hahn, U. (2016). Bad company - neighborhoods in neural embedding spaces considered harmful. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2785–2796.

Herbelot, A. and Baroni, M. (2017). High-risk learning: acquiring new word vectors from tiny data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 304–309.

Hill, F., Reichart, R., and Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Hinrichs, E., Hinrichs, M., Kübler, S., and Trippel, T. (2019). Language technology for digital humanities: introduction to the special issue. *Language Resources and Evaluation*, 53(4):559–563, Dec.

Kabbach, A., Gulordava, K., and Herbelot, A. (2019). Towards incremental learning of word embeddings using context informativeness. In *Proceedings of the 57th Conference of the Association for Computational Linguistics: Student Research Workshop*, pages 162–168, Florence, Italy, July. Association for Computational Linguistics.

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

Lazaridou, A., Marelli, M., and Baroni, M. (2017). Multimodal word meaning induction from minimal exposure to natural text. *Cognitive science*, 41:677–705.

Manjavacas, E., Long, B., and Kestemont, M. (2019). On the feasibility of automated detection of allusive text reuse. In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage,*

Social Sciences, Humanities and Literature, pages 104–114.

McGillivray, B. and Kilgarriff, A. (2013). Tools for historical corpus research, and a corpus of Latin. *New Methods in Historical Corpus Linguistics*, (3):247–257.

Meyer, F., Oortwijn, Y., Sommerauer, P., Bloem, J., Betti, A., and Fokkens, A. (2019). The semantics of meaning: distributional approaches for studying philosophical text. Unpublished manuscript.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of the ICLR Workshop*.

Nivre, J., De Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., et al. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Passarotti, M. (2019). The project of the Index Thomisticus Treebank. In Monica Berti, editor, *Digital Classical Philology. Ancient Greek and Latin in the Digital Revolution*, volume 10 of *Age of Access? Grundfragen der Informationsgesellschaft*, pages 299–319. De Gruyter.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA. `http://is.muni.cz/publication/884893/en`.

Reynaert, M., Bos, P., and van der Zwaan, J. (2019). Granularity versus Dispersion in the Dutch Diachronical Database of Lexical Frequencies TICCLAT. In *Proceedings of CLARIN Annual Conference 2019 – Conference Proceedings*, pages 169–172, Leipzig, Germany. CLARIN ERIC.

Reynaert, M. (2010). Character confusion versus focus word-based correction of spelling and OCR variants in corpora. *International Journal on Document Analysis and Recognition*, pages 1–15. 10.1007/s10032-010-0133-5.

Sahlgren, M. and Lenci, A. (2016). The effects of data size and frequency range on distributional semantic models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 975–980.

Schick, T. and Schütze, H. (2019). Rare words: A major problem for contextualized embeddings and how to fix it by attentive mimicking. *arXiv preprint arXiv:1904.06707*.

Schnabel, T., Labutov, I., Mimno, D., and Joachims, T. (2015). Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 298–307.

Sprugnoli, R. and Passarotti, M., (2020). *EvaLatin 2020 Shared Task Guidelines*. Version 1.0, December 10, 2019.

Sprugnoli, R., Passarotti, M., and Moretti, G. (2019). Vir is to Moderatus as Mulier is to Intemperans: Lemma embeddings for Latin. In *Sixth Italian Conference on Computational Linguistics*, pages 1–7.

Turney, P. D. and Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188.

Van den Berg, H., Betti, A., Oortwijn, Y., Parisi, M. C., Wang, S., and Koopman, R. (ongoing). The Spread of the Mathematical Method in Eighteenth-Century Germany: A Quantitative Investigation.

van Gompel, M., van der Sloot, K., Reynaert, M., and van den Bosch, A. (2017). FoLiA in practice: The infrastructure of a linguistic annotation format. In J. Odijk et al., editors, *CLARIN-NL in the Low Countries*, chapter 6, pages 71–81. Ubiquity (Open Access).

van Wierst, P., Hofstede, S., Oortwijn, Y., Castermans, T., Koopman, R., Wang, S., Westenberg, M. A., and Betti, A. (2018). Bolvis: visualization for text-based research in philosophy. In *3rd Workshop on Visualization for the Digital Humanities*.

Zeman, D., Hajic, J., Popel, M., Potthast, M., Straka, M., Ginter, F., Nivre, J., and Petrov, S. (2018). CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual parsing from raw text to universal dependencies*, pages 1–21.