# A computational approach to the study of semantic change in Latin: the case of Christian Latin vocabulary

Valentina Lunardi

University of California, Los Angeles

5th September 2024

## Overview

- This paper builds on a study of the *Itinerarium Egeriae* lexicon, previously presented at LVLT XIV (Lunardi 2022).
- Initial focus: explore Christian Latin vocabulary, using the *Itinerarium Egeriae* as a starting point.
- Initial aim: quantify the impact of Christianity on Latin lexicon and its influence on the development of the Romance languages.
- Challenges: close-reading methods are time-consuming, requiring extensive reading of texts from pre-Christian possibly to modern times.
- Shift to computational and quantitative approaches to handle large amounts of data efficiently.
- Today:
    - Preliminary results of study using static word embeddings to investigate meaning changes triggered by the spread of Christianity.
    - Focus on two items selected from the *Itinerarium*.
    - Compare word embedding results with close-reading analyses.
    - Study currently restricted to Latin, ending at 600 CE.
- Goals:
    - Scale up the study of Christian vocabulary across texts and centuries.
    - Contribute to understanding Christian Latin and the debate on a Christian register.
    - Advance the study of semantic change in historical linguistics.

## Roadmap

1 Word embeddings for semantic change: overview

2 Analysis

  2.1 Close-reading

  2.2 Corpus design

  2.3 Word embedding model and parameters

  2.4 Word embeddings results

3 Conclusions

## 1 Word embeddings for semantic change: overview

- The fundamental principles behind word embeddings are:
  - The distributional hypothesis: "The degree of semantic similarity between two linguistic expressions A and B is a function of the similarity of the linguistic contexts in which A and B can appear" (Lenci 2008, 3).
  - Vector representation of words (Jurafsky and Martin 2023, 106–7).
- For a more practical example, consider the following set of three documents:
  1. "The **horse** jumped on the bed."
  2. "The **horse** teased the unicorn."
  3. "The unicorn jumped on the bed."
- To find a vector for 'horse' with with a window size of 1 for co-occurrence, we count the number of times it appears next to each of the words present in the three documents. The words are 'the', 'horse', 'jumped', 'on', 'bed', 'teased', 'unicorn'.
- 'Horse' appears next to 'the' twice, never next to 'horse', once next to 'jumped', never next to 'on', never next to 'bed', once next to 'teased', never next to 'unicorn' – the coordinate values for 'horse' are $(2, 0, 1, 0, 0, 1, 0)$.
- 'Horse' is a vector in a 7-dimensional space, each dimension corresponding to the words 'the', 'horse', 'jumped', 'on', 'bed', 'teased', 'unicorn'.
- This is a simplified example of a word vector, but the same principles lie behind the more advanced neural network-based models (such as word2vec, fasttext, BERT, etc.).
- Word vectors produced by these models are known as word embeddings, and these models need large corpora to produce them.
- The process through which these models learn the coordinate values is called "training". This also involves setting certain parameters to guide the process, such as vector size, minimum frequency, window size for co-occurrence, etc.
- The spatial representation of words allows us to quantify the difference between two words by comparing their vectors: this is done via cosine similarity, whose value ranges from $0$ to $1$ (with $0$ indicating no similarity and $1$ indicating maximum similarity) (Jurafsky and Martin 2023, 112–3, McGillivray and Tóth 2020, 66–7).
- Cosine similarity can be used to detect semantic change via quantification of the difference between

two vectors of the same word, where the two are calculated from two separate sets of documents from different timeframes.

- A useful related feature that these models can provide is a list of the word embeddings which are "closest" (within the same corpus) to the vector of the word we are interested in: these are known as an embedding's "neighbours".

## 2 Analysis

- Close-reading analysis of two semantically-changed items selected from the *Itinerarium*.
- Details of computational test, including corpus design, an outline of some of the parameters used for the word embedding model, and a discussion of the results for the same two items to compare with the close-reading analysis.

### 2.1 Close-reading

#### 2.1.1 *deus*

- Attested 101 times in the *Itinerarium Egeriae*, very frequent in my working corpus more generally.
- With the advent of Christianity, *deus* began to be used to refer to the Christian god in addition to referring to the Roman gods (Gudeman 1912).
- Its high frequency is a strong motivation behind this choice, as it will most likely reflect in a high-quality embedding.

#### 2.1.2 *commūnicō*

- Attested 8 times in the text, always with the meaning 'to receive the Holy Communion' (see e.g. III, 6; XVI, 7)
- In Classical Latin it can mean 'to share / take a share in (something with someone)', 'to impart / communicate (information or knowledge)', 'to discuss (something) together with (someone)', and it is only transitive (Bannier 1911).
- Starting with Tertullian we get an intransitive use with the meaning 'to participate' (Bannier 1911).
- In the *Vetus Latina*, the verbs which intransitive *commūnicō* translates are either κοινωνέω (e.g. in Romans 12:13, 1 Timothy 5:22, 1 Peter 4:13), συγκοινωνέω (e.g. in Ephesians 5:11, Philippians 4:14, Revelation 18:4), or μετέχω (e.g. in 1 Corinthians 10:21).

- *commūnicō* seems to have undergone a process through which Latin lexemes either acquire a new sense or their range of meaning is restricted prompted by the Greek lexemes they are translating (see Burton 2000, 120–8).
- Looking back at the use of *commūnicō* in the *Itinerarium Egeriae*, we find that the verb is also intransitive there (in all eight cases), so the meaning 'to receive the Holy Communion' seems to be a narrowing of the intransitive use found starting in Tertullian.

### 2.2 Corpus design

- LatinISE is a corpus of Latin conceived by McGillivray and Kilgariff (2013) containing approximately 13 million words.
- The corpus size is reduced to use texts from 300 BCE to 600 CE, for a total of 5 million words:
  - The end date depends on the willingness to research Latin while it was a living language.
  - The start date was chosen to make the pre- and post-Christian subcorpora chronologically balanced, with the split coinciding with the first attestations of Christian texts.
- Two chronologically-determined subcorpora, with the split coinciding with the first attestations of Christian texts, currently set to 150 CE.
  - These allow for comparison of embeddings for the same words across the two timeframes, the first of which should show no influence from Christianity.
- Two genre-determined subcorpora contained in the second timeframe, one containing exclusively Christian texts, the other all non-Christian ones.
  - These allow for comparison of embeddings for the same words across different sets of texts within the same timeframe.
- The subcorpora are fairly balanced with their counterparts in terms of number of tokens.

### 2.3 Word embedding model and parameters

- I used the code by McGillivray (2023) as a starting point.
- In the original code, the choice of model and the values for various parameters conform to the findings of Sprugnoli, Passarotti, and Moretti (2019), Sprugnoli, Moretti, and Passarotti (2020), and Ribary and McGillivray (2020).
- The model used is fastText, given its use of *n*-grams (i.e., subwords) during training, making it particularly suitable for morphologically rich languages.

| First timeframe<br>(frequency: 2746) | | Second timeframe<br>(frequency: 11974) | | Christian subcorpus<br>(frequency: 11248) | | Non-Christian subcorpus<br>(frequency: 726) | |
|---|---|---|---|---|---|---|---|
| dea | 0.584 | creator | 0.571 | gloria | 0.597 | sanctus | 0.781 |
| Iuppiter | 0.527 | dominus | 0.530 | pater | 0.584 | gaudium | 0.751 |
| numen | 0.500 | Christus | 0.520 | iudico | 0.566 | auris | 0.741 |
| divinus | 0.491 | peccator | 0.508 | maledico | 0.564 | inquiam | 0.733 |
| religio | 0.473 | divinitas | 0.506 | glorifico | 0.560 | pietas | 0.721 |
| expio | 0.464 | glorifico | 0.504 | confiteor | 0.559 | inquam | 0.709 |
| caelestis | 0.453 | angelus | 0.481 | Christus | 0.547 | mens | 0.705 |
| consecro | 0.451 | visibilis | 0.480 | propterea | 0.537 | gratus | 0.702 |
| fas | 0.444 | salvator | 0.479 | peccator | 0.536 | animus | 0.702 |
| invoco | 0.432 | dilectio | 0.475 | dilectio | 0.536 | oro | 0.692 |

Table 1: Neighbors for *deus*

- A lemmatised corpus is used to reduce variability given the already limited size of the corpus.

- Working on my own adaptation of the code, I found that couple of parameters were especially relevant for my corpus and results:

  – `min_count`, the parameter that regulates the minimum frequency required for a word to be included in the training;

  – and `max_n`, an optional parameter which, if set to 0, allows for the exclusion of subwords.

- It is common to exclude low-frequency words from training, as the embeddings generated for these words often end up being unrepresentative. However, for this kind of corpus with many low-frequency items, `min_count` is especially important. We will see this more concretely when discussing the individual embeddings.

- Using `max_n` to exclude subwords defeated the purpose of using fastText, but it was necessary to address the model's tendency to favor orthographically similar words.

- For example, without this adjustment, among the neighbours for *deus*, we would not only find *semideus* 'demigod', but also *lapideus* 'stony'.

## 2.4 Results

### 2.4.1 *deus*

- With `min_count` = 50, I obtained the neighbors presented in table 1.

- Between the first and second timeframe, there is a visible difference in terms of neighbors, as there are some clear associations with:

| First timeframe (frequency: 70) | | Second timeframe (frequency: 109) | |
|---|---|---|---|
| laudo | 0.533 | praesumo | 0.701 |
| probo | 0.529 | consentio | 0.609 |
| penes | 0.523 | agnosco | 0.575 |
| delibero | 0.521 | prosum | 0.551 |
| suadeo | 0.520 | accommodo | 0.549 |
| absens | 0.512 | respuo | 0.540 |
| amicus | 0.509 | displiceo | 0.540 |
| adsentio | 0.504 | praescribo | 0.540 |
| praesertim | 0.502 | noceo | 0.539 |
| commemoro | 0.499 | imputo | 0.535 |

Table 2: Neighbors for *commūnicō* – attempt 1

  - Roman religion in the first timeframe (e.g. *dea* implies the existence of goddesses and therefore polytheism, *Iuppiter* is the chief deity of the Roman pantheon), and
  - Christian religion in the second timeframe (e.g. *creātor* and *dominus*, both frequent designations for the Christian god).

- The neighbors in the Christian subcorpus confirm association with Christian religion.
- By contrast, a few of those in the non-Christian subcorpus are less representative of concepts pertaining to god and religion (e.g. *auris*, *inquiam*, *inquam*).
- This may be due to the considerably lower amount of tokens for *deus* in the non-Christian subcorpus: this a first proof of the fact that higher word frequency means higher-quality embeddings.
- If we leave aside these odd items, there are others that could be connected with religion, without specific connections with Christianity.
- The value of cosine similarity for *deus* is 0.610 between the first and the second timeframe, and 0.462 between the Christian and non-Christian subcorpora.

### 2.4.2 *commūnicō*

- Training with the value for `min_count` only yielded results for the two chronologically-determined subcorpora, while for the genre-determined subcorpora fastText was unable to find an embedding for it due to low-frequency.
- The results from the first and second timeframe are presented in table 2.

| First timeframe<br>(frequency: 70) | | Second timeframe<br>(frequency: 109) | | Christian subcorpus<br>(frequency: 93) | | Non-Christian subcorpus<br>(frequency: 16) | |
|---|---|---|---|---|---|---|---|
| delibero | 0.597 | praesumo | 0.739 | deprehendo | 0.944 | subsero | 0.966 |
| absens | 0.594 | consentio | 0.710 | retineo | 0.933 | immineo | 0.965 |
| praesertim | 0.567 | recognosco | 0.659 | comparo | 0.926 | superesset | 0.964 |
| commemoro | 0.563 | subicio | 0.657 | competo | 0.921 | hortulanus | 0.962 |
| laudo | 0.560 | agnosco | 0.656 | liceo | 0.921 | exitialis | 0.961 |
| conservo | 0.558 | exhibeo | 0.650 | reprehendo | 0.920 | contemplor | 0.960 |
| mereo | 0.557 | denego | 0.633 | profiteor | 0.919 | spolio | 0.960 |
| comparo | 0.555 | deputo | 0.630 | cedo | 0.919 | insidiarum | 0.959 |
| ignosco | 0.554 | approbo | 0.630 | deputo | 0.918 | punitur | 0.959 |
| mereor | 0.543 | accommodo | 0.627 | quodsi | 0.912 | pertimesco | 0.959 |

Table 3: Neighbors for *commūnicō* – attempt 2

- Despite the fact that fasttext seems to struggle a bit more to yield representative embeddings for low-frequency words, a few of the neighbors for *commūnicō* are close to its semantic field.

- For the first timeframe, *dēlīberō* can mean ‘to take counsel’ or ‘to advise upon’, and one of the possible meanings of *commemorō* is ‘to make mention of something’; neither of these are far off the Classical Latin meaning of *commūnicō*, ‘to share (something with someone)’ and ‘to discuss (something) together with (someone)’.

- For the second timeframe, only *cōnsentiō* stands out as being close in meaning to *commūnicō*: it can mean ‘to determine in common’, ‘to agree’, which is fairly close to the ‘to share (something with someone)’ meaning of *commūnicō*, with the implication that agreeing on something means sharing an opinion.

- Let us lower the minimum frequency threshold (set `min_count` to 5) and look first at the neighbors for *commūnicō* with this changed parameter, presented in table 3.

- Embeddings are produced for all subcorpora this time.

- For the first timeframe, the two words which we identified as closer to *commūnicō* are now closer to the top of the list of neighbors.

- Among the neighbors from the second timeframe:

  - *cōnsentiō* has higher cosine similarity than it did before (from 0.61 to 0.71);

  - *exhibeō* can mean ‘to show’, which again is not far in meaning from ‘to share(something with someone)’.

- As for the Christian subcorpus, *competō* (‘to come together’) and *profiteor* (‘to declare publicly’ or ‘to show’) can similarly be linked to the ‘to share(something with someone)’ meaning of *commūnicō*.
- However, there is no clear connection to the more specific meaning of ’to receive the Holy Communion’.
- Neighbours for the non-Christian subcorpus, by contrast, do not seem very good.
- For *commūnicō*, the different parameters gave us some slightly better embeddings.

## 3 Conclusions

- We can conclude that:
  - The results are promising for high-frequency words such as *deus*, as embeddings seem to represent them with good accuracy.
  - The lower the frequency of the word in the corpus, however, the less satisfying the results.
  - Embeddings can help us greatly where a word is very common in a corpus, but philological analysis is still extremely valuable for low-frequency words.
  - Yet, these methods proved extremely powerful for common words, promising to aid our study of semantic change significantly.
- As next steps for improvement of these results, I plan to use contextual embeddings and increase the size of the corpus.

## References

Bannier, Wilhelm. 1911. "commūnico, -āvī, -ātum, -āre." In *Thesaurus Linguae Latinae Online,* 3:1954–1960. Berlin, New York: De Gruyter. Available at https://tll.degruyter.com/article/3_0_9_communico_v2007.

Burton, Philip. 2000. *The Old Latin Gospels: A Study of their Texts and Language.* Oxford: OUP.

Gudeman, Alfred. 1912. "deus, -eī m. dea, -ae f." In *Thesaurus Linguae Latinae Online,* 5:885–915. 1. Berlin, New York: De Gruyter. Available at https://tll.degruyter.com/article/5_1_04_deus_2_v2007.

Jurafsky, Dan, and James H. Martin. 2023. "Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition." Available at https://web.stanford.edu/~jurafsky/slp3/.

Lenci, Alessandro. 2008. "Distributional semantics in linguistic and cognitive research." *Rivista di Linguistica,* no. 20 (1): 1–31.

Lunardi, Valentina. 2022. *The Role of Christian Latin in the Development of Romance Vocabulary: A Lexical Study of the Peregrinatio Egeriae.* Available at https://valentinalunardi.com/talks/ghent_peregrinatio.

McGillivray, Barbara. 2023. *Semantic change in LatinISE.* Available at https://github.com/BarbaraMcG/latinise/blob/master/lvlt22/Semantic_change.ipynb, accessed 2023-09-01.

McGillivray, Barbara, and Adam Kilgariff. 2013. "Tools for historical corpus research, and a corpus of Latin." In *New methods in historical corpus linguistics,* edited by Paul Bennett, Martin Durrell, Silke Scheible, and Richard J. Whitt, 247–257. Tübingen: Narr.

McGillivray, Barbara, and Gábor Mihály Tóth. 2020. *Applying Language Technology in Humanities Research: Design, Application, and the Underlying Logic.* Palgrave Macmillan.

Ribary, Marton, and Barbara McGillivray. 2020. "A Corpus Approach to Roman Law Based on Justinian's Digest." *Informatics* 7 (4).

Sprugnoli, Rachele, Giovanni Moretti, and Marco Passarotti. 2020. "Building and Comparing Lemma Embeddings for Latin. Classical Latin versus Thomas Aquinas." *Italian Journal of Computational Linguistics* 6 (1): 29–45.

Sprugnoli, Rachele, Marco Passarotti, and Giovanni Moretti. 2019. "Vir is to Moderatus as Mulier is to Intemperans - Lemma Embeddings for Latin." In *Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019,* edited by Raffaella Bernardi, Roberto Navigli, and Giovanni Semeraro, vol. 2481. CEUR Workshop Proceedings. CEUR-WS.org.