UCLA PROGRAM IN INDO-EUROPEAN STUDIES

The impact of Christianity on the Latin lexicon: computational approaches to semantic change

Valentina Lunardi
vlunardi@g.ucla.edu

Semantic change

WHAT IS IT

- Change in the meaning of words over time
- OE *mete* ‘food’ > PDE *meat*
  - “narrowing”
- Lat. *salārium* ‘(soldier’s) allotment of salt’ > ‘(soldier’s) wages’ > ‘wages’
  - “widening”
- All languages have evidence of this phenomenon

Semantic change

WHAT IS IT

- A significant portion of the scholarship on semantic change has been devoted to the classification of its various types:
  - Narrowing vs. widening
  - Amelioration (e.g. Lat. *caballus* ‘nag, workhorse’ > It. *cavallo* ‘horse’) vs. pejoration (e.g. Lat. *sinister* ‘left’ > Sp. *siniestro* ‘sinister’)
  - Metaphor (e.g. Lat. *testa* ‘pot’ > It. *testa* ‘head’)
  - Metonymy (e.g. Lat. *māxilla* ‘jaw’ > Sp. *mejilla* ‘cheek’),
  - Hyperbole (e.g. Lat. *mȳthicus* ‘mythical, pertaining to myths’ > It. *mitico* ‘legendary, awesome’)
  - Etc.

Semantic change

STATUS QUO ANTE

- For a long time, the “black sheep” of historical linguistics:

  “Any attempt at a systematic study of semantic change, in fact, will yield only limited rewards, for two reasons: with rare (and not very helpful, however interesting) exceptions, ***semantic change is completely patternless***; and, second, insight is forestalled by our nearly perfect ignorance of the real nature of the semantic component of language.”

  Sihler 2000: 94 (emphasis mine)

- Why?

Semantic change

STATUS QUO ANTE

1) The motivations behind semantic change:
   - Numerous, varied, and often extra-linguistic (e.g. dependent on cognitive factors or sociocultural forces)
2) The method to trace semantic change:
   - The only available one (until recently) was to follow the meaning of each individual word within a language (or language branch) across time through close-reading of texts
   - Each language has tens of thousands of words, making it difficult to find trends of change (and therefore predictability) with this method

Semantic change

WHAT HAS CHANGED

- Over the past couple of decades, the field of semantic change has seen some promising advancements
- On one hand, linguistic-oriented literature (esp. advanced by Traugott) has focused on changes pertaining to function words and has identified several unidirectional patterns of change, e.g. via:
  - Subjectification
    - *very*: 'true, real' > intensifier (from objective to subjective evaluation)
  - Grammaticalization
    - *will*: 'to want' > future marker (from content word to grammatical marker)

Semantic change

WHAT HAS CHANGED

- On the other hand, further advancements are made possible by the increasing availability of digital textual corpora
- Computationally-oriented literature has adopted quantitative, statistical, and machine-learning methods to analyze the now easily accessible textual data
- Such research has included the development of methods to detect semantic change within diachronic corpora (see e.g. Tahmasebi et al. 2021)
- Today we will look at one such method

Semantic change

WHAT HAS CHANGED

- Crucial observation:
  - Hundreds of texts can be processed with a computer in a matter of minutes/hours/days (depending on the task) – way faster than a human being ever could
- It now seems achievable to try and find ***trends of change***, not just for function words but also for content words
- Not my goal per se, as this kind of conclusion would require comparing change across languages to form conclusions about typological trends
- I can, however, contribute to this in the form of data about trends of change for a specific language

Christian Latin

BASICS

- The spread of Christianity had an effect on the politics, society, culture, and unsurprisingly also the languages of the western world
- That Christianity had influence on Latin (and consequently on the development of Romance) is generally unchallenged
- The *Paradebeispiel* in this sense comes from Löfstedt (1959: 81)
  - *parabola* ‘parable’ and *parabolāre* ‘to tell in a parable’ respectively came to mean ‘word’ and ‘to speak’ in some Romance languages
  - they ended up substituting *verbum* ‘word’ and *loquor* ‘to speak’, relatively basic Classical Latin lexical items

Christian Latin

BASICS

- Most phenomena are a result of influence from Greek:
  - Loanwords: *angelus* ‘angel, herald of God’ ⟵ ἄγγελος
  - Loan translations: *glōrificō* ‘glorify’ ⟵ δοξάζω
  - Sematic loans: *virtūs* ‘miracle’ ⟵ δύναμις

Burton 2011: 489

- Some previously infrequent syntactic devices used by Christian writers are sometimes argued to result from their frequency in (literal) biblical translations

Christian Latin

THE DEBATE

- However, “Christian Latin” as a distinct linguistic entity is debated
- The most influential and controversial work on Christian Latin was produced within the Nijmegen school
- Their theory is known as the *Sondersprache* hypothesis, which has its canonical statement in Schrijnen (1932) and is extensively developed by Mohrmann (see esp. Mohrmann 1958–1977)
- It states that the Christian community developed a unique form of Latin, a register distinct from non-Christian Latin variants in its morphology, syntax, and lexicon

Christian Latin

THE DEBATE

- The *Sondersprache* hypothesis has always faced criticism (see e.g. Marouzeau 1932, Coleman 1987), especially about the following:
  - Overvaluing of the extent of communal living among early Latin-speaking Christians and exaggeration of the impact this would have had on their language
  - Heavy reliance on evidence from a limited number of educated writers, when Christian writers often differ in style from each other as much as their pagan peers

Christian Latin

THE DEBATE

- Christian Latin is made up of

> “several distinct registers: the vulgarised Latin of Bible and Psalter, the plain but unvulgarised style of ecclesiastical administration, the more sophisticated idiom of expository and hortatory literature and finally the products of high literary culture – the hymns and collects of the Liturgy and Offices”
>
> Coleman 1987: 52

Christian Latin

HOW TO MOVE FORWARD

- Even the harshest critics of the *Sondersprache* hypothesis acknowledged the existence of a unique Christian vocabulary which distinguishes it from other sociolects of Latin (Burton 2011: 487–8)
- Its signs of influence from the Greek scriptures possibly attracted the most attention
- A study of Christian Latin vocabulary and its interaction with the Latin lexicon more generally might improve our understanding of Christian Latin as a whole

Ultimate goals

- Leverage the new methods to contribute to:
  - The issue of semantic change
  - The study of the history of Latin and its development into the various Romance languages
  - The debate about Christian Latin

Plan for (the rest of) today

- Choose lexemes for analysis + outline meaning change
- Briefly introduce a computational/quantitative method:
  - Static word embeddings
- Outline corpus and subcorpora for analysis
- Evaluate results by comparing with information gathered through close-reading and/or consultation of lexicographical resources
- Conclude + mention other methods (current/future work)

Selected lexemes

HOW/WHAT

- Words for analysis generally picked through either reading of primary texts or secondary literature
- (Originally started my word collection by reading the *Itinerarium Egeriae* a few years back)
- A mix of high frequency but more obvious words and more interesting but less frequent words
- Today: 2 words total for illustrative purposes

Selected lexemes

COMMŪNICŌ

- Attested 8 times in the *Itinerarium*, 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)

Selected lexemes

COMMŪNICŌ

- In the *Vetus Latina*, intransitive *commūnicō* translates:
  - κοινωνέω (e.g. Romans 12:13, 1 Timothy 5:22, 1 Peter 4:13),
  - συγκοινωνέω (e.g. Ephesians 5:11, Philippians 4:14)
  - μετέχω (e.g. 1 Corinthians 10:21, although in patristic quotations (Cyprian, Jerome))
- *commūnicō* seems to have been the vehicle of semantic loan ('semantic extension' in Burton 2000: 120–8)
- The use of *commūnicō* in the *Itinerarium Egeriae* is also intransitive: the meaning 'to receive the Holy Communion' seems to be a narrowing of the intransitive use found starting in Tertullian
- (But cf. Mod. Greek κοινωνώ, which can also have this meaning)

Selected lexemes

DEUS

- Very frequent in working corpus
- 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 reflect in a high-quality representation for the static embeddings method

Introduction to word embeddings

FUNDAMENTAL IDEAS

- Distributional semantics: within theoretical Linguistics, popular in the 1950s, later supplanted by formal semantics
  - 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)

Introduction to word embeddings

Table 1 Co-occurrence matrix of the words *strawberry, blackberry, blueberry,* and *computer* with the context words *all, eat, nice, office, phone, raspberry, Samsung, software* in the BNC Spoken corpus

	all	eat	nice	office	phone	raspberry	Samsung	software
strawberry	9	6	5	0	0	5	0	0
blackberry	0	0	3	0	3	5	4	0
blueberry	0	5	4	0	0	7	0	0
computer	0	0	0	3	12	0	0	4

McGillivray, B. (2022, Jul 12). How to Use Word Embeddings for Natural Language Processing. SAGE Publications Ltd. https://doi.org/10.4135/9781529609578

Introduction to word embeddings

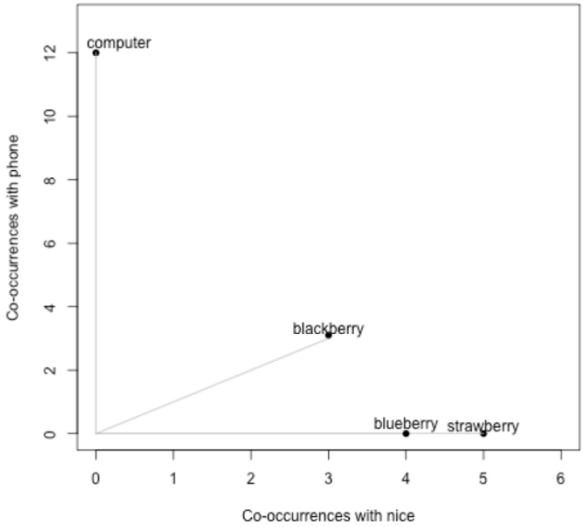

Figure 1. Vector representation of the words *strawberry*, *blackberry*, *blueberry,* and *computer* in a bi-dimensional space.

The horizontal axis corresponds to the concurrence with *nice*, and the vertical axis corresponds to co-occurrences with *phone*. *Strawberry* will have the coordinates (5,0), *blackberry* (3,3), *blueberry* (4,0), and *computer* (0,12).

McGillivray, B. (2022, Jul 12). How to Use Word Embeddings for Natural Language Processing. SAGE Publications Ltd. https://doi.org/10.4135/9781529609578

Introduction to word embeddings

PRACTICAL EXAMPLE

- 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’.

Introduction to word embeddings

PRACTICAL EXAMPLE

- ‘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 therefore (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.)

Introduction to word embeddings

THE NEWER MODELS

- Only the vectors produced by the newer models are called “word embeddings”
- These newer models still use vectors to represent the meaning of words, but the crucial difference is that their coordinate values are found automatically
- This is done through a process called “training”, which also relies on the distribution of words in the training corpus and is subject to certain pre-set parameters, such as window size for co-occurrence, minimum frequency of a word for inclusion in training, etc.

Introduction to word embeddings

THE NEWER MODELS

- For earlier models, each value represents a frequency count, but in the newer models these values are not easily interpreted
- One way to think about it is that each value is representative of some property of the word, e.g.
  - ‘unicorn’: high value for the coordinate representing its fictitiousness, high value for its horse-like qualities, low value for its inanimacy
- In practice, however, these properties are not defined in a way that is interpretable by humans
- Words that appear in similar contexts will have similar values

Introduction to word embeddings

COSINE SIMILARITY

- The spatial representation of words allows us to quantify the difference between two words by comparing their vectors
- Unit: cosine similarity, with value ranging from 0 to 1
  - 0 indicating no similarity, 1 indicating maximum similarity (Jurafsky and Martin 2023, 112–3)
- How can we use cosine similarity to detect semantic change?
  - Compare vectors of the same word, where the vectors are calculated from (two or more) separate sets of documents from different timeframes

Introduction to word embeddings

COSINE SIMILARITY

- 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” and can be e.g. synonyms or words within the same semantic field
  - Neighbours are also given a cosine similarity score, essentially allowing us to see which words appear in the most similar environments to the one under scrutiny

Corpus design

DEFINING THE TIME FRAME

- LatinISE is a corpus of Latin conceived by McGillivray and Kilgarriff (2013) containing approximately 13 million words
- The corpus size is reduced to use texts from 300 BCE to 600 CE, for a total of 6.8 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

Corpus design

DEFINING THE SUBCORPORA

- Two chronological subcorpora, with the split coinciding with the first attestations of Christian texts, currently set to 150 CE
  - These allow for comparison of representations for the same words across the two timeframes, the first of which should show no influence from Christianity
- Two subcorpora contained in the second timeframe, one containing exclusively Christian texts, the other all non-Christian ones
  - These allow for comparison of representations 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

Model setup

- Starting point: code by McGillivray (2023)
- Choice of model and values for parameters conforming largely to the findings of Sprugnoli, Passarotti, and Moretti (2019), Sprugnoli, Moretti, and Passarotti (2020), and Ribary and McGillivray (2020)
- Lemmatised corpus to reduce variability
- fastText picked for suitability for morphologically rich languages, given its use of n-grams (i.e. subwords) during training
- Some tested parameters were:
  - Window size: both 5 and 10 tested
  - Minimum frequency: 50 for high freq., 5 for low freq. words
  - Exclusion / inclusion of subwords during training

Results

DEUS: NEIGHBOURS

300 BCE - 150 CE (freq. 5408)		150 – 600 CE (freq. 15086)		Christian (freq. 13928)		Non-Christian (freq. 1158)	
immortalis	0.723	*dominus*	0.712	*pater*	0.622	*numen*	0.856
numen	0.718	*omnipotens*	0.661	*omnipotens*	0.599	*sanctus*	0.842
superi	0.694	*Christus*	0.625	*dominus*	0.590	*Christus*	0.831
dea	0.656	*creator*	0.622	*maiestas*	0.586	*pietas*	0.817
Iuppiter	0.652	*iustitia*	0.611	*creator*	0.579	*pius*	0.815
propitius	0.598	*factor*	0.586	*exalto*	0.569	*o*	0.780

Results

DEUS

- Between the first and second timeframe, there is a visible difference in terms of neighbours, with clear associations with:
  - Roman religion in the first
  - Christian religion in the second
- Cosine similarity: 0.666
- Within the Christian subcorpus the associations with Christianity are confirmed
- Within the non-Christian one, we have a mix, signaling that the change in meaning of *deus* affected the language as a whole

Results

COMMŪNICŌ: NEIGHBOURS

300 BCE - 150 CE (74)		150 – 600 CE (106)		Christian (95)		Non-Christian (11)
praesertim	0.607	*praesumo*	0.536	*deprehendo*	0.876	–
Chrysogonus	0.607	*presbyterium*	0.528	*praesumo*	0.857	–
vitupero	0.554	*ieiuno*	0.528	*praesumptio*	0.854	–
discepto	0.549	*alteruter*	0.522	*adscribo*	0.842	–
separatim	0.548	*inquino*	0.522	*recognosco*	0.838	–
collega	0.548	*clericus*	0.521	*consentio*	0.837	–

Results

COMMŪNICŌ

- Admittedly, these results are not very good
- Some backstory:
  - LatinISE put together from two major sources, one with poetry only and one with mostly prose
  - Recently improved by removing clashing duplicates
  - But… the results I got before adding in the data from the poetry-only source were slightly more representative of the word
  - Corpus approx. 5 million tokens before addition of poetry-only source
  - frequency of *commūnicō* virtually unchanged
- Let's look at previous results!

Results

COMMŪNICŌ: NEIGHBOURS TAKE 2 (LITTLE POETRY)

300 BCE - 150 CE (70)		150 – 600 CE (109)		Christian (93)		Non-Christian (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	*supersum*	0.964
commemoro	0.563	*subicio*	0.657	*competo*	0.921	*hortulanus*	0.962
laudo	0.560	*agnosco*	0.656	*licet*	0.921	*exitialis*	0.961
conservo	0.558	*exhibeo*	0.650	*reprehendo*	0.920	*contemplor*	0.960

Results

COMMŪNICŌ

- These are better (thought perhaps not yet good) results
- How genre of training corpus affects results:
  - Something for me to think about!!!
- Only for a few of these neighbours we can try to argue for similarity
- For the first time slice:
  - *dēlīberō* can mean ‘to take counsel’, ‘to advise upon’
  - *commemorō* can mean ‘to make mention of something’
  - not far off the ‘to share (something with someone)’ and ‘to discuss (something) together with (someone)’ meaning of *commūnicō*

Results

COMMŪNICŌ

- For the second time slice:
  - *cōnsentiō* can mean ‘to determine in common’, ‘to agree’
  - *exhibeō* can mean ‘to hold forth’, ‘to show’
  - both close to the ‘to share (something with someone)’ meaning of *commūnicō*

- For the Christian subcorpus:
  - *competō* ‘to come together’
  - can similarly be linked to the same meaning of *commūnicō*

Results

COMMŪNICŌ

- There is definitely no clear connection to the more specific meaning ‘to receive the Holy Communion’
- Can we see a shift towards intransitivity? Hard to say
- Neighbours for the non-Christian subcorpus, by contrast, are really quite unsalvageable…

Conclusions

- The results are promising for high-frequency words such as *deus*
  - embeddings seem to represent them with pretty good accuracy!!
- The lower the frequency of the word in the corpus, however, the less satisfying the results
  - See next slide for some alternative (non-close-reading) solutions
- (Static) embeddings can help us greatly where words are very common in a corpus, making our attempts to trace changes in meaning much less time-consuming
- *Of course*, philological analysis is still extremely valuable, and this is especially true for low-frequency words
- Yet, this method promises to aid our study of semantic change significantly

Next steps

- Try out different methods!!
  - Collocational analysis
  - Contextual embeddings
- Compare and contrast to assess which is more effective and/or whether a mix of approaches is needed
- Think more about how the genres present in corpus affects results
- But also assess current shortcomings
  - We want to advance the field by making our job easier
  - But, of course, we want the final product to be good!

Some essential bibliography

Coleman, Robert. 1987. “Vulgar Latin and the diversity of Christian Latin.” In *Latin vulgaire – Latin tardif: Actes du 1er Colloque International sur le latin vulgaire et tardif, Pécs, 2-5 septembre 1985*, edited by József Hermann, 37–52. Tübringen: Niemeyer.

Mohrmann, Christine. 1958–1977. *Études sur le latin des chrétiens. 4 vols*. Rome: Storia e letteratura.

Sihler, Andrew L. 2000. *Language History : An Introduction*. Amsterdam: John Benjamins Publishing.

Schrijnen, Jos. 1932. *Charakteristik des altchristlichen Latein*. Nijmegen: Dekker en Van de Vegt & Van Leeuwen.

Burton, Philip. 2011. “Christian Latin.” In *A Companion to the Latin Language*, edited by James Clackson, 485–501. Oxford: Wiley-Blackwell.

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

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/.

Tahmasebi, Nina, Lars **Borin**, and Adam **Jatowt**. 2021. “Survey of Computational Approaches to Diachronic Conceptual Change.” In *Computational approaches to semantic change*, edited by Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu, and Simon Hengchen, 1–91. Berlin: Language Science Press.

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.

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

Thank You

Personal website: