

Obtaining More Expressive Corpus Distributions for Standardized Ancient Languages

Oliver Hellwig^{a,b}, Sven Sellmer^{a,c} and Sebastian Nehrdich^{a,d}

^a*Institute for Language and Information, Heinrich Heine Universität, Düsseldorf*

^b*Department of Comparative Language Science, University of Zürich*

^c*Institute for Oriental Studies, Adam Mickiewicz University, Poznań*

^d*Khyentse Center for Tibetan Buddhist Textual Scholarship, Universität Hamburg*

Abstract

This paper introduces a latent variable model for ancient languages that aims at quantifying the influence that early authoritative works exert on their literary successors in terms of lexis. The model jointly estimates the amount of word reuse, based on uni- and bigrams of words, and the date of composition of each text. We apply the model to a corpus of pre-Renaissance Latin texts composed between the 3rd c. BC and the 14th c. CE. Our evaluation focusses on the structures of word reuse detected by the model, its temporal predictions and the quality of the inferred diachronic distributions of words, which last aspect is assessed using a newly designed task from the field of computational etymology.

Keywords

Text reuse, citations, standardized languages, historical corpora, Bayesian mixture model

1. Introduction

Constructing diachronic trajectories of word¹ frequencies seems to pose no major technical challenges. Given a database of timestamped texts and their linguistic annotations, one can derive such trajectories by applying smoothing techniques (e.g. temporal binning, kernel-based techniques) to the frequencies of words in individual texts. In the fields of Historical Linguistics and Classical Studies matters can, however, become more complicated because word frequencies can be influenced by various confounding factors such as the dialect or mother tongue spoken by an author, by changes in the orthography, or by language standardization, on which we focus in this paper. Following the definition given by Joseph [1], we use the term ‘standardized language’ for a codified, prestigious language variety that is mainly used for administrative and literary purposes. Examples of such languages include Latin as used in the post-Classical period or Sanskrit in the form prescribed by the grammarian Pāṇini.

While the vocabularies, as well as stylistic features, of standardized languages may still change (see e.g. Clackson [2] for Latin and Wackernagel [3, XXIIff.] for Sanskrit), phonetics and morpho-syntax remain, so to say, frozen or undergo only minor diachronic changes. The work described in this paper primarily addresses the question to which degree the word usage in


CHR 2021: Computational Humanities Research Conference, November 17–19, 2021, Amsterdam, The Netherlands

✉ Oliver.Hellwig@uni-duesseldorf.de (O. Hellwig); sellmer@hhu.de (S. Sellmer); nehrdich@uni-duesseldorf.de (S. Nehrdich)

🆔 0000-0003-0387-2827 (O. Hellwig); 0000-0002-6688-0667 (S. Sellmer); 0000-0001-8728-0751 (S. Nehrdich)

© 2021 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

¹The term ‘word’ denotes the lemma of a word throughout this paper if not specified otherwise.

standardized languages reflects the everyday language use of authors who often spoke ‘vulgar’ varieties of the standardized language or, later, vernacular languages stemming from these varieties.

Two factors are especially relevant here. First, many authors writing in standardized languages show the tendency to reuse and paraphrase authoritative works which were considered as a kind of gold standard (see e.g. Lee [4]; Roberts [5] for Latin). The influence of earlier works can therefore bias and distort the distributions of words found in later ones. More generally, such languages typically are conservative in that they preserve words that are no longer current outside literary circles. An instructive example for such a trend is the Latin word *equus* ‘horse’ [see 6, 291]. While this word is the standard expression for ‘horse’ in Classical Latin and does not have any archaic ring to it, the Romance languages, which originated from Latin dialects spoken in the late Antiquity (“Vulgar Latin”, see Herman [7]), derive their words for ‘horse’ from Latin *caballus* (e.g. Fr. *cheval*, It. *cavallo*), which suggests that occurrences of *equus* in post-Classical Latin texts no longer reflect the spoken language.

Second, the temporal structure of ancient corpora may be (partly) unclear, making it even more difficult to reliably construct diachronic lexical trajectories. The accumulated effects that standardization, text reuse, semantic conservatism and temporal uncertainties exert on corpus distributions are difficult to determine from the raw corpus data alone, making it necessary to balance the corpus evidence with detailed qualitative – and time-consuming – studies of individual words. Such issues are not restricted to Latin texts of the late Antiquity and the Middle Ages [see e.g. 8], but are also found, for example, in Buddhist Chinese [see e.g. 9], in the Indic corpora composed in Sanskrit and Pāli, or in Classical Chinese [10]. As these languages are the ancestors of important modern language families, Classical Studies as well as Linguistics can benefit from corpus distributions that distinguish between the actual language use and the influence of authoritative works.²

This paper discusses a Bayesian mixture model for lemmatized texts that disentangles the influence exerted by authoritative, frequently cited and paraphrased texts on the word usage encountered in their literary successors. It aims at generating a clearer picture of the actual practice in standardized languages, at quantifying the amount of word reuse and at unveiling intellectual lineages in such corpora. For modelling word reuse, this paper builds on previous research that quantifies the influence of cited authors in the context of scientific publications [11, 12]. Unlike such bibliometric studies, citations in ancient corpora are mostly not (clearly) marked as such and must therefore be inferred from the data in the approach presented in this paper. The detection of literary influences can be further enhanced by inspecting lexical n-grams. While many previous approaches represent the textual data as bags of words, one may argue that text reuse and stylistic influences rather get manifest in collocations taken over from earlier literary works. While the presence of the unigrams *aurum* ‘gold’ and *pretiosus* ‘precious’ only gives a weak indication of literary ancestry, a bigram formed of these two words (in *pretiosior auro* ‘more precious than gold’) is a much clearer indication that the late Roman author Maximianus has been influenced by the Augustan poet Ovid. Our model therefore complements the bag of words representation with lexical bigrams [see 13] and makes the decision for uni- or bigrams part of the inference process.

Another important aspect is the time of composition. Most (Bayesian) mixture models with

²The expression ‘actual language use’ has to be taken in a technical sense that changes according to the author: For authors speaking some form of Latin, it refers to the language they use in everyday situations; for users of other languages, it denotes, somewhat artificially, the Latin they write, but from which the effects of word reuse have been removed, so to speak.

a temporal component assume that the time of composition is an observed variable (e.g. Blei and Lafferty [14], Wang et al. [15]). Such an assumption does not hold for many ancient texts as their dates are either unknown or still under scrutiny. While there exist some Latin texts whose dates of composition are strongly disputed (see e.g. Laurioux [16] on the cookbook of Apicius), this problem is more urgent for ancient Indian corpora, where dates proposed for early texts are often just educated guesses (see e.g. Olivelle [17], 7-13 on the Sanskrit philosophical texts called Upaniṣads). We address this issue by modelling the time of composition of each text as a latent variable that conditions the observed features and incorporates the current state of scholarly research with the help of a temporal prior (see Hellwig [18] for a related approach for Vedic Sanskrit).

We use Latin texts composed between the 3rd c. BCE and the 14th c. CE as a test case. As Sec. 5 will show, many aspects of the evaluation rely on qualitative arguments, as gold standards for these tasks are currently not available. Using the Latin corpus offers the advantage that the evaluation can build on a long history of literary and linguistic research, so that our results can be compared against an extensive record of previous scholarship. The initial application to the well-researched Latin tradition makes it easier to transfer the methods developed here to more disputed textual traditions of South Asia.

After an overview of related work in Computational Linguistics (Sec. 2), Sections 3 and 4 describe the data and the model. Section 5 assesses various choices in the model design using posterior predictive checks (Sec. 5.1) and presents an evaluation of three prominent aspects of our model: word reuse (Sec. 5.2), predicted times (Sec. 5.3) and the inferred corpus distributions (Sec. 5.4), the latter being tested on a new task in computational etymology. – Data and scripts are available at <https://github.com/OliverHellwig/sanskrit/tree/master/papers/chr2021>.

2. Related research

Our model of word reuse builds on previous work on detecting citation activities in scientific literature. Such activities have repeatedly been formalized using (ad-)mixture models, starting with Cohn and Hofmann [19] whose generative model conditions citations on the presence of hidden topics. Erosheva et al. [20] extend Latent Dirichlet Allocation by conditioning the generation of links on the same document-specific topic distributions as the generation of words. The citation-influence model of Dietz et al. [11], also assuming citations to be fully observed, splits the process of generating words in two branches: a word in document d is either drawn from the topic distribution of a cited text (which is in turn sampled from a document-specific multinomial distribution over citable documents) or from a word distribution specific to d (“innovation”). Nallapati et al. [12] present two models that treat citations as latent variables sampled on the basis of document-specific topic distributions. Although not directly concerned with citations, the author-topic model of Rosen-Zvi et al. [21] offers an alternative view of what we want to achieve in this paper as some texts in ancient standardized languages can indeed be considered the work of a collective of – not necessarily contemporaneous – authors (see e.g. Colledge [22] on the composition of the *Legenda aurea* by an anonymous group of authors).

Previous research has proposed various admixture models that contain a temporal component modelled either in discrete bins (e.g. Blei and Lafferty [14] or Frermann and Lapata [23] with Gaussian priors on logistic topic-word mixtures) or as continuous observed variables (e.g. Wang and McCallum [24]; Wang et al. [15]). More complex models as e.g. proposed by

Table 1

Composition of the corpus. The first column gives the historical period according to Adamik [31] (also see Sec. 5.3).

Period	Authors	Tokens
Old	3	9,653
Classical	59	2,630,289
Late	46	1,765,834
Transitional	13	90,656
Medieval	45	823,974

Kawamae [25] split the generation of words in time- and document-specific branches.

Using bigrams in admixture models was first proposed by Wallach (2006; also see Nokel and Loukachevitch 2016 for a survey). While Wallach models all data points as bigrams, the collocation model of Griffiths et al. [28] makes the decision for uni- vs. bigrams part of the model structure. Wang et al. [13] further make the decision for uni- vs. bigrams dependent from the hidden topic.

3. Data

The experiments described in this paper are based on the works of 166 Latin authors who were active between the 3rd c. BCE and the 14th c. CE, the French philosopher Nicole Oresme (1320-1382) being the latest one included. From among the available Latin corpora (for an overview see McGillivray [29, ch. 2]), we chose the Latin library corpus of the CLTK library³ due to its wide coverage. An author is included if at least 50k of text are contained in the CLTK library or if the author is considered important for (text-)historical reasons (e.g. the *Res gestae* of Augustus). The raw source data are unbalanced (authors such as Cicero or Thomas Aquinas are strongly over-represented), and individual works are often split into multiple files. We therefore merge all works of one author into a single text, although, arguably, the preference for citing and reusing text can vary inside the oeuvre of an author.

Latin is a strongly inflectional language. In addition, the orthography of some source texts has not been standardized, and especially the late Christian authors are responsible for some variation so that working with raw textual data would result in very sparse feature matrices. All texts are therefore lemmatized using Collatinus [30] (which manages to resolve many of the non-standard spellings in the process) and these lemmatized versions constitute the data used for all following steps of the processing pipeline. After removing 104 stop words such as *ad* ‘to(wards)’, *et* ‘and’ or *meus* ‘my’ as well as lemmata that occur less than 30 times, the corpus consists of 5,320,406 word tokens with 10,309 distinct lemmata (also see the summary in Tab. 1). Public sources such as the Encyclopedia Britannica and Wikipedia are used for gathering information about the lifetime of each author (l_d, u_d : birth and death years of author d). If not specified otherwise, the date m_d of a text d denotes the mean of this time span, i.e. $m_d = \frac{1}{2}(l_d + u_d)$.

³thelatinlibrary.com, <http://cltk.org/>

4. Model

The model discussed in this paper needs to deal with three types of uncertainty: (1) unknown structures of word reuse; (2) fuzzy or unknown dates of composition; (3) the question whether uni- or bigrams of words should be used as the observed features. This leads to the following generative story (see eq. 2 for the complete specification): First, for the i th word in text d , the source text c_{di} is drawn from a text-specific multinomial distribution ξ_d . Note that ξ_d includes the text d itself. Such self-loops mean that the respective data point is peculiar to the actual author of text d .⁴ While many citation models proposed so far can build on a given citation structure (as e.g. defined by web links or scholarly citations in articles), this information is not available for our data. The value of the prior α_{ij} (text i cites from text j) therefore needs to be adapted during inference depending on the inferred latent times. After each iteration of the Gibbs sampler (this means after running it once over all data points), the mean time slots μ of all texts are calculated based on the current state of the latent temporal assignments, and the value of α_{ij} is updated using a sigmoid function:

$$\alpha_{ij} = \begin{cases} 10 & \text{if } i = j \\ 0 & \text{if } \mu_j - \mu_i > 3 \\ \frac{1}{1 + \exp(-(\mu_i - \mu_j))} & \text{else} \end{cases} \quad (1)$$

The high value for α_{ii} encourages the model to explain the words observed in a text by the preferences of its author. Note that the zeros for the case $\mu_j - \mu_i > 3$ are structural zeros so that text j is not considered a possible source of i if $\alpha_{ij} = 0$.⁵ In addition, we multiply each element of α with a *citation mask* $m \in \{0, 1\}^{D \times D}$ that is derived from running a Levenshtein-based citation detector over the unlemmatized texts. The value m_{ij} is set to 1 if at least one sequence of five or more words is shared by texts i, j ; else to zero. Zero values in m are again interpreted as structural zeros. The use of this mask is based on the idea that literal citations, as detected by the Levenshtein algorithm, indicate the acquaintance of an author with a previous work and thus increase the probability that individual words from this previous work are used as well.

Second, a time slot t_{di} is drawn from a text-specific multinomial temporal distribution $\omega_{c_{di}}$. The prior $\beta_{c_{di}}$ of $\omega_{c_{di}}$ incorporates the current state of scholarly knowledge about the time of composition of text c_{di} , and possible time slots obtain a flat uniform prior in the range $l_{c_{di}}, u_{c_{di}}$ while slots outside $[l_{c_{di}}, u_{c_{di}}]$ are set to structural zeros.

Third, the model draws a Bernoulli-distributed variable b_{di} that decides if the word x_{di} and its successor x_{di+1} typically form a bigram. Contrary to the model proposed by Wang et al. [13], this decision does not depend on the sampled time t_{di} and thus saves $(T - 1) \cdot V^2$ trainable parameters. Based on the sampled value of b_{di} , either the unigram x_{di} or the bigram $x_{di}x_{di+1}$ is drawn from time-specific multinomial distributions $\phi_{t_{di}}^U$ resp. $\phi_{t_{di}x_{di}}^B$.

With Θ denoting all trainable parameters and π all priors, the joint distribution is given by

⁴This choice is represented by the Beta distributed variable λ in Dietz et al. [11].

⁵The difference of three time slots is motivated by the following idea: As will be shown in Sec. 5.1, 150 is a good choice for the number of time slots. As the whole corpus covers a temporal range of about 1,700 years, three time slots correspond to slightly more than 30 years, a span that may describe the active period of one author.

Table 2

Variables, dim(ensions), par(ameters), c(ounters) and pr(iors) of a model with D documents, T time slots and a vocabulary size of V

Variable	Dim.	Par.	C.	Pr.
text \rightarrow citation	$\mathbb{R}^{D \times D}$	ξ	\mathbf{A}	α
citation \rightarrow time	$\mathbb{R}^{D \times T}$	ω	\mathbf{B}	β
time \rightarrow unigrams	$\mathbb{R}^{T \times V}$	ϕ^U	\mathbf{C}^U	γ^U
time \rightarrow bigrams	$\mathbb{R}^{T \times V \times V}$	ϕ^B	\mathbf{C}^B	γ^B
2 words \rightarrow uni-/bigr.	$\mathbb{R}^{V \times V}$	ψ	\mathbf{L}	δ

the following equation (notation in Tab. 2):

$$\begin{aligned}
p(\mathbf{c}, \mathbf{b}, \mathbf{t}, \mathbf{x}, \boldsymbol{\Theta} | \boldsymbol{\pi}) &= \prod_d^D \text{Dir}(\xi_d | \alpha_d) \prod_d^D \text{Dir}(\omega_d | \beta_d) \prod_u^T \text{Dir}(\phi_u^U | \gamma^U) \\
&\cdot \prod_u^T \prod_v^V \text{Dir}(\phi_{uv}^B | \gamma^B) \prod_v^V \prod_w^V \text{Beta}(\psi_{vw} | \delta) \\
&\cdot \left[\prod_d^D \prod_i^{n_d} [\text{Cat}(c_{di} | \xi_d) \text{Cat}(t_{di} | \omega_{c_{di}}) \cdot \text{Bern}(b_{di} | \psi_{x_{di}x_{di+1}}) (b_{di} \text{Cat}(x_{di+1} | \phi_{t_{di}x_{di}}^B) \right. \\
&\quad \left. + (1 - b_{di})(\text{Cat}(x_{di} | \phi_{t_{di}}^U))] \right] \tag{2}
\end{aligned}$$

The blocked Rao-Blackwellized Gibbs Sampler [32] is obtained by using Dirichlet-multinomial integrals:

$$\begin{aligned}
p(c_{di} = e, t_{di} = k, b_{di} = l, x_{di} = u, x_{di+1} = v | \mathbf{c}^{-di}, \mathbf{t}^{-di}, \mathbf{x}^{-di}, \boldsymbol{\Theta}, \boldsymbol{\pi}) \\
\propto (\mathbf{A}_{de}^{-di} + \alpha_{de}) \frac{\mathbf{B}_{ek}^{-di} + \beta_{ek}}{\sum_l^T \mathbf{B}_{el}^{-di} + \beta_{el}} \begin{cases} (\mathbf{L}_{x_{di}x_{di+1}}^{1(-di)} + \delta^1) \frac{\mathbf{C}_{kuv}^{B(-di)} + \gamma_v^B}{\sum_w^V \mathbf{C}_{kvw}^{B(-di)} + \gamma_w^B} & b_{di} = 1 \\ (\mathbf{L}_{x_{di}x_{di+1}}^{0(-n)} + \delta^0) \frac{\mathbf{C}_{ku}^{U(-di)} + \gamma_u^U}{\sum_w^V \mathbf{C}_{kw}^{U(-di)} + \gamma_w^U} & b_{di} = 0 \end{cases}
\end{aligned}$$

A small, but important difference to models that operate with a known citation structure is the selection of possible sources. In this paper, a text c is only considered as a possible source for an observed uni- x_{di} or bigram $x_{di}x_{di+1}$ if it also contains x_{di} or $x_{di}x_{di+1}$. This condition prevents the model from assigning too much weight to early authors such as Cicero and Vergil.

5. Experiments

This section reports qualitative and quantitative evaluations for the three relevant elements of our model: the detected structure of word reuse (Sec. 5.2), the temporal predictions (Sec. 5.3) and the diachronic trajectories of words that can be inferred from it (Sec. 5.4).

5.1. Architecture and Parameter Settings

We use posterior predictive checks (PPC; Mimno et al. [33]) to compare various model architectures and parameter settings. Given a trained model, we draw textwise samples of the observed words using Eq. 2 and compare these samples with the true distributions in each text using

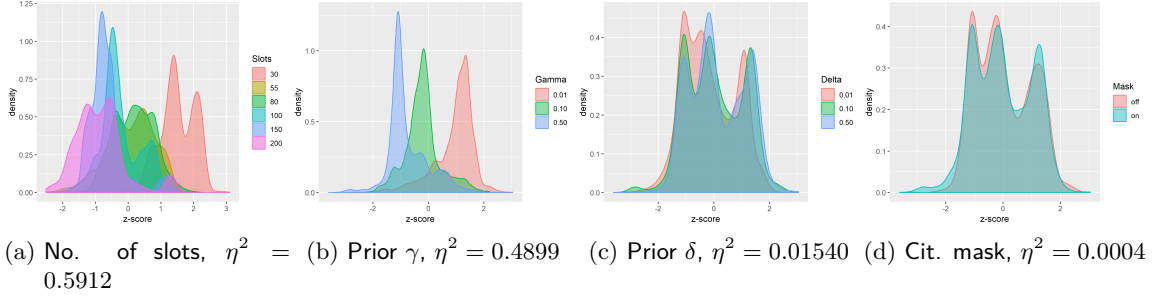


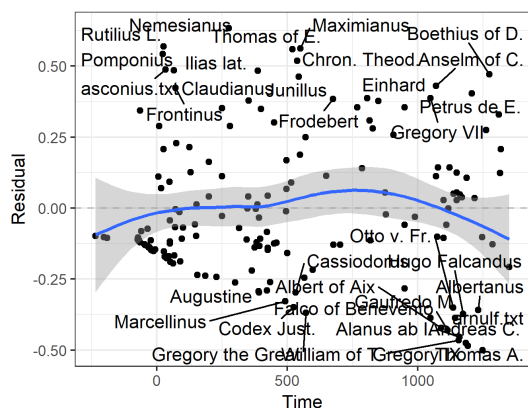
Figure 1: Results of the posterior predictive checks and Cohen’s η^2 . Each colored curve shows the density of the z-standardized values for one setting. Small z-scores are better.

the Hellinger Distance. The values that result from 30 replications per text are grouped by texts and z-standardized, and ANOVAs are performed in order to test for significant differences between settings. Figure 1 shows smoothed density estimates of these z-scores for four central design choices: the number of temporal slots (Fig. 1a), the parameters γ (time \rightarrow feature; Fig. 1b) and δ (uni- or bigram; Fig. 1c) and the use of the precomputed citation mask (Fig. 1d). While ANOVA points to (highly) significant differences in all four settings, the values of Cohen’s η^2 which quantify the effect size and are displayed below each subfigure indicate that only the number of slots and the prior γ have a relevant influence on the outcome of the model, while the influence of δ and the citation mask must be considered as very small. Based on this evaluation, we choose 150 time slots, $\gamma = 0.5$, $\delta = 0.01$ for all following experiments, and we apply the citation mask.

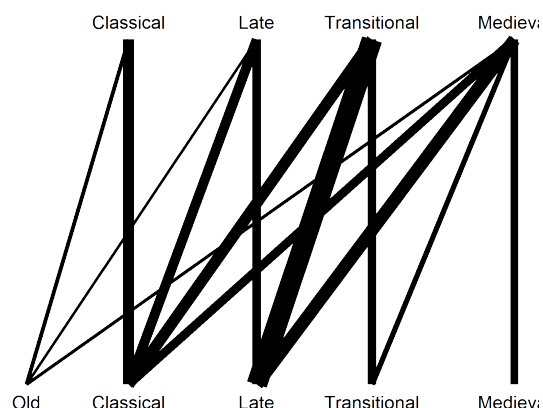
Running another PPC for establishing the optimal number of iterations of the Gibbs sampler, we found no significant differences between models trained with 100, 300, 500 or 1,000 iterations (p-value of the ANOVA: 0.163). This somehow unexpected result is certainly due to the fact that our model already has rather strong priors induced by the structural zeros in the citation mask and the temporal prior β so that only few iterations are required to obtain a good representation of the data. We therefore run the sampler for 100 iterations and record the sampled values once after the last iteration.

5.2. Word Reuse

As mentioned in the introduction, understanding the intellectual lineages of historical corpora is one important aim of this paper. Therefore, the evaluation starts with inspecting the inferred structure of word reuse. We calculate, for each text d , the proportion of words labeled as reused, i.e. for which $c_{di} \neq d$ according to the model output. These proportions can be expected to be correlated with the true date of d , as later texts have more opportunities to reuse words than earlier ones. In order to deal with this effect, we perform a partial correlation by fitting a linear regression that predicts the proportions of words labeled as reused (y) based on the number of possible source texts (x). The residuals of this regression, which capture how much the model output deviates from the linear estimate, are plotted against the true date of each text (see Fig. 2a). Here, the dashed horizontal line at $y = 0$ corresponds to a residual of 0 and thus to a perfect prediction of the model output through the linear regression. The blue curved line is a smoothed density estimate of the actual residuals. This smoothed estimate shows that the proportions of word reuse conform to the values estimated by the linear model until the end of



(a) Residuals of a linear regression that predicts the inferred number of reused words given the number of available source authors. Individual authors are labeled if their residuals fall in the 5% resp. 95% quantiles.



(b) Schematical representation of word reuse, grouped by literary periods. The source periods are found at the bottom. The line width indicates the strength of the activity.

Figure 2: Patterns of word reuse detected by the model

the Late Antiquity (5th c. CE). We observe increasing word reuse in the 8th or 9th c. CE, a period commonly known as the Carolingian Renaissance, which saw a revival of classical Latin literature that accompanied the formation of the Carolingian state [see e.g. 34]. In the 10th c. CE and later, the proportions of word reuse tend to fall below their expected values. Seen from the perspective of literary history, the intensive word reuse in the Carolingian Renaissance is connected with authors such as Hrabanus Maurus, Angilbert or Alcuin, who in his *De rhetorica* freely mixes extracts from Cicero’s *De inventione* and other authoritative sources with his own comments [see e.g. 35]. In the 10th c., a new form of Latin is constituted, which, though still accepting the classical language as its gold standard, is strongly influenced by the idiom of Christian theological authors (“Ecclesiastical Latin”, see e.g. Dinkova-Bruun [6]). This form of medieval Latin can therefore be expected to share less lexical features with Classical Latin than earlier forms of the language.

Figure 2b presents another view of the literary influences. In this plot all texts are aggregated by the five literary periods defined by Adamik [31], plus an extra period “Medieval Latin” starting at 900 CE.⁶ The widths of the lines between target, i.e. “citing” (top), and source, i.e. “cited” (bottom), periods indicate the relative amount of word reuse inferred by the model. The plot shows that works from the classical era quite constantly remained important sources of word reuse throughout all periods considered in this paper, although even their influence begins to wane in the Transitional Period (600–900 CE) and the Middle Ages. Such a result makes sense as the works of some classical authors did not survive the breaks in the political and religious history and were only rediscovered in the Italian Renaissance or even later (see e.g. Tutrone [36] on the limited reception of the important Roman philosopher Lucretius in the (early) Middle Ages). The strong connections between Late Latin on one hand and the transitional and medieval periods on the other are due to the numerous important Christian texts composed in Late Antiquity, most notably the Latin translation of the Bible (*Vulgata*)

⁶We label the period called Vulgar Latin by Adamik as Late Latin in this paper in order to distinguish it from the sub-standard variety discussed by Herman [7].

and the work of Augustine. In addition, Fig. 2b shows a decline in word reuse between the Transitional Period, which comprises the Carolingian Renaissance just discussed, and Medieval Latin – most authors from the Transitional Period were obviously not too much regarded in later times.

In order to understand which authors are mainly responsible for the distribution observed in Fig. 2b, we collect, for each literary period, those three authors with the highest amount of words marked as reused, applying a minimal threshold of 1,000. The resulting list contains the following authors:

Old Cato (the Elder) is the only representative of old Roman literature, a result which is in accordance with his extraordinary importance for the development of a genuinely Latin literature. His compendia on agriculture and warfare as well as the collection of his orations (compiled by himself) exerted a considerable influence on later authors [37, 340-41].

Classical Ovid and Cicero can be seen as the top representatives of Latin poetry and prose, while Livy stands for the genre of classical historiography.

Late This period shows an interesting interference between the famous Christian author Augustine and the Vulgata, a new translation of the Bible composed by Jerome. Different from what may be expected, Augustine is more frequently marked as cited than the Vulgata (161,501 vs. 74,445 times). A closer inspection of words and bigrams labeled as cited reveals that the model has problems in assigning individual Biblical citations to the Vulgata or the Vetus Latina, the older Latin version of the Bible preferably cited by Augustine [see e.g. 38, 36-39]. – The third representative of this period is Gregory of Tours, best known for his historical writings.

Transitional Here, only Beda has made it in the list – a result fully in accordance with his popularity in the Middle Ages [39].

Medieval While Thomas Aquinas is a central representative of medieval Latin and its focus on theological discussions, Albert of Aix and William of Tyre represent the genre of medieval historical writings with a special focus on the Crusades.

To sum up this section, it appears that the model was able to recover structures of word reuse that conform to scholarly expectations.

5.3. Timestamping

We model the partly unclear times of composition as latent variables. In this section we assess the quality of the resulting temporal predictions. We simulate a research setting in which only approximate temporal information is available, by setting the temporal ranges of all D texts d to the ranges of the literary periods containing them according to Adamik [31, 9]. These artificially obfuscated ranges are used as temporal priors β_d (see eq. 2). All texts are trained jointly, and we evaluate how well the model can recover the exact dates and the correct temporal order of the texts. Notably, this experiment is not merely another academic exercise, but bears practical implications when studying ancient Indian text corpora for which only approximate temporal information is available [see 18]. – Table 3 reports two evaluation measures:

- The period-wise mean absolute error (MAE) calculated as $\frac{1}{|\{d \in P\}|} \sum_{d \in P} \|m_d - \mu_d\|_1$ where μ_d is the mean of the word-wise temporal assignments for text d , and P is the literary period.
- Ranking accuracy: The texts are grouped by their literary periods, and all texts belong-

Table 3

Grouped mean absolute errors (MAE; in years) and ranking accuracies of the temporal predictions for five literary periods

Period	MAE	Rank acc.
Old	41.1	0.0
Classical	87.7	52.5
Vulgar	101.3	48.6
Transitional	67.4	59.0
Medieval	136.5	40.3

ing to one period are ordered by their true dates m_d . The ranking accuracy gives the proportion of text pairs for which the predicted temporal order is the same as the true one.

The results in Tab. 3 show that dating texts composed in standardized languages is challenging. Although the literary periods only extend over 200-300 years each, the MAEs vary between 40 and 140 years and thus cover substantial parts of each period. It may, however, be noted that Kumar et al. [40] report slightly higher MAEs of 85-155 years for English stories published between 1798 and 2008, which suggests that the results achieved by our model are actually in an acceptable range. The values of the ranking accuracy are coupled with the uncertainties in the temporal predictions and fall below the random baseline of 50% for three of the five periods. Notably, both evaluation measures seem to get worse for post-Classical texts when Latin gradually ceased to be used as a spoken language, and an ANOVA of the MAEs as well as a Fisher-Yates test of the raw counts for the ranking accuracies both show (highly) significant differences between all periods (p-values: 0.00147 [MAE]; 0.0005 [ranking acc.]).

In order to assess if the temporal predictions improve when more reliable temporal information is available, we perform a cross-validation experiment. A subset of fifteen authors⁷ is chosen as the test set. For each text in this set, we obfuscate its date in the same way as in the first experiment, while all $D - 1$ other texts keep their temporal gold information. The model is trained with the $D - 1$ training texts for 100 iterations and then for another 100 iterations with the combined training and test set (see the method `Gibbs1` in Yao et al. [41]). The results are compared with the predictions made by the Topic over Time model [24] which is often used as a baseline for latent variable models with a temporal component.⁸ The results in Tab. 4 show that our model is slightly, but not significantly better than ToT (p-value of a paired directed Wilcoxon test: 0.26). While ToT occasionally assigns all texts from one period to the same date range, our model better captures the temporal dynamics. This impression is confirmed when calculating the ranking accuracy (ours: 60%; ToT: 33%) for the data in Tab. 4.

⁷This limitation is due to time constraints. We choose three authors from the start, middle and end of each period; see the first column of Tab. 4.

⁸We use 150 topics and all hyperparameter settings as described in the original paper. The predicted time slot is that with the highest posterior $\arg\max_t \sum_i^{n_d} \log p(t|\psi_{z_i})$, with the additional constraint that $l_d \leq t \leq u_d$, in order to make a fair comparison with the model presented in this paper; see Sec. 2 in Wang and McCallum [24].

Table 4

Cross-validated temporal predictions of the model in this paper and ToT [24]. The best prediction per text is printed bold. Predictions that fall in the true temporal range of a text are underlined.

Text	Date	This paper	ToT
Naevius	-270/-201	-179	-298
Ennius	-239/-169	-229	-298
Cato	-234/-149	-260	-298
Cicero	-106/-43	78	-6
Seneca Y.	-4/65	120	-4
Apuleius	123/170	121	1
Commodianus	225/275	407	374
Leo the Great	390/461	431	<u>406</u>
Maximianus	500/600	413	374
Chron. Fredegar	600/700	824	813
Alcuin	735/804	868	735
Erchempert	850/900	678	753
Leo of N.	900/1000	1095	1312
Bernard de C.	1100/1200	1138	1300
Nicole O.	1320/1382	1297	1291
MAE		93.2	114.8

5.4. Features

Getting a more realistic picture of how words are diachronically distributed in standardized languages is an important aim of this paper. This section therefore compares the linguistic expressiveness of empirical corpus distributions with those inferred by our model. Using posterior estimates of the variational parameters (i.e. $\omega'_{dt} = \frac{B_{dt} + \tau_{dt}}{\sum_u B_{du} + \tau_{du}}$ etc.) based on those cases in which the model assigns words to unigrams, we obtain the conditional probabilities $p(x|d)$ of a word x given a text d by marginalizing the latent citations and temporal assignments:

$$p(x|c) = \sum_c^D \sum_t^T p(c|d)p(t|c)p(x|t) = \sum_c^D \sum_t^T \xi'_{dc} \omega'_{ct} \phi^{U'}_{tx} \quad (3)$$

We expect that the diachronic trajectories of this conditional distribution differ from the corpus distribution of a word x when the use of x in later texts is mainly due to literary influences.

In order to quantitatively support the claim that the inferred distributions yield a more realistic description of the actual language use, we address the problem of predicting lexical stability [see 42]. While substantial parts of the vocabulary of Romance languages can be derived from precursors in (Vulgar) Latin by applying rules of regular sound change [43, 42], there are important individual words such as *equus* ‘horse’ or whole classes of words such as the vocabulary of war that do not have derivatives in the Romance languages. Apart from various socio-cultural factors (on which see e.g. Campbell [44], 244ff. and especially Vincent [45] on Latin vocabulary that was “submerged” in classical works), the frequency of use in the spoken language is a determining factor for the survival or obsolescence of a word [46]. If the inferred distributions better capture the actual use than the corpus distributions, they should better be able to predict the survival of Latin words in the Romance languages.⁹

⁹A factor we ignored as non-essential for the present purpose, but which should be taken into account in

As the etymological information in Wiktionary is incomplete and noisy [47], we collect all Latin words that are recorded as etyma of Romance words in Meyer-Lübke [48], a standard reference work of Romance etymologies. Although scholarly research has revised some decisions made in this work, it is still considered as a largely complete collection of surviving Latin etyma (see e.g. Stefenelli [42], 568) so that words not recorded there can be assumed not to have derivatives in Romance languages. From among the 10,308 words in our vocabulary, 2691, i.e. 26.1%, have such a derivative.¹⁰

We aggregate the empirical and inferred distributions by various ranges of years, z-standardize these binned values and use them as input features for a feed-forward neural network with four hidden units and softplus activations. The neural network is trained on the binary prediction task whether or not a Latin word has derivatives in any Romance language.¹¹

Figure 3a shows the F-scores (y-axis) depending on the sizes of the temporal bins applied (x-axis). While the F-scores generally decrease with increasing sizes of the temporal bins, the F-score of the inferred distributions is consistently higher than that of the empirical ones. The drop of the F-score is especially obvious when using the empirical distributions with a bin size of 30 years instead of the unbinned distributions (“all”). The failure of the model that uses the empirical distribution is due to its low recall in these cases. In order to better understand the behaviour of the predictor, we collect all inherited words that were labelled correctly using the inferred, but wrongly using the empirical distributions, calculate their empirical and inferred distributions and smooth these distributions with a Gaussian kernel. Figure 3c contrasts the means (plus/minus one standard deviation) of the two groups. The plot shows that the inferred distribution transfers probability mass from occurrences in (late) classical texts to the (early) Middle Ages (\sim 8th c.+), i.e. to a period in which the Romance languages are generally assumed to develop. A similar effect can be observed for words which are only predicted correctly when using the empirical distributions (see Fig. 3d). Apparently, the mixture model has missed effects of word reuse in these cases, as it assigns too much weight to occurrences in the early Middle Ages. Finally, when examining distributions of inherited words detected by neither classifier, it becomes apparent that many of them are popular in classical and medieval texts, but rare in the Late Antiquity and the Transitional Period (see e.g. the plots for *expecto* ‘expect’ in Fig. 3b). Although the mixture model draws up the distributions for the critical phase of the early Middle Ages, this effect is not strong enough to make the classifier label such words as inherited.

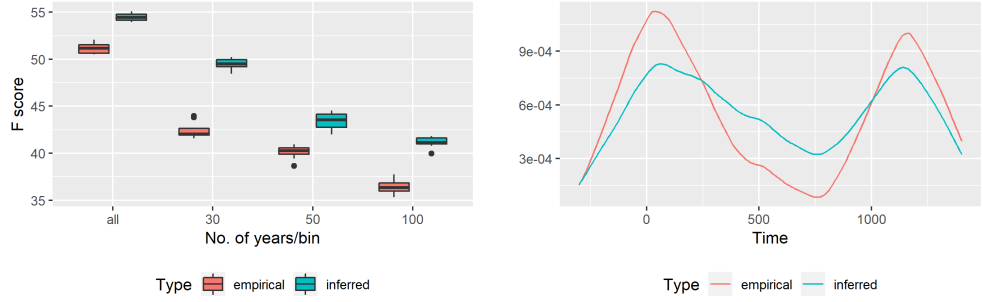
6. Summary

Diachronic corpora are indispensable tools for studying linguistic developments and intellectual lineages in premodern societies. Depending on the degree of standardization which the corpus language has undergone as well as on the amount of text reuse, linguistic distributions extracted from diachronic corpora can be misleading because the language usage of authori-

future, more detailed studies, is the fact that, starting from the early Middle Ages, for a growing number of authors their mother tongue is not a Romance language but belongs to another family (mostly the Germanic one).

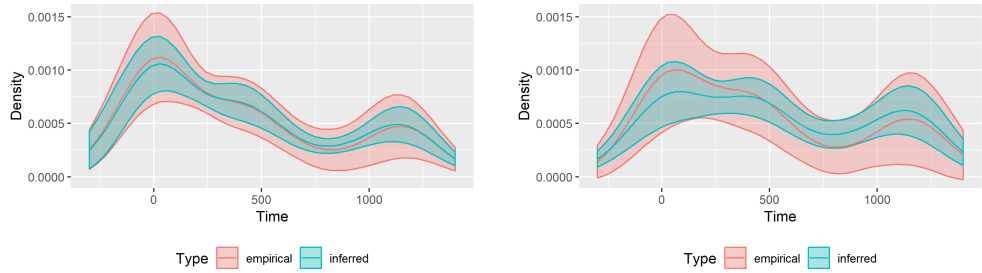
¹⁰Note that this number only covers Romance words derived by regular sound change, but not, for example, borrowed words.

¹¹Meyer-Lübke [48] does not consistently report all Romance derivatives of a given Latin word, so that we could not formulate this problem as a multi-class prediction task. – Apart from a simple neural network, we also tested flat ML models such as logistic regression, but found our approach to perform better.



(a) F-scores of predicting the lexical stability of Latin words in Romance languages; values reported without grouping ("all") and grouped by the number of years per temporal bin.

(b) Normalized and smoothed empirical and inferred distributions of the word *expecto* 'expect'



(c) Inferred correct, empirical wrong

(d) Inferred wrong, empirical correct

Figure 3: Results of the etymology prediction task: F-scores when using the empirical resp. inferred distributions (Fig. 3a); smoothed accumulated distributions for cases in which only the inferred (Fig. 3c) or empirical distributions (Fig. 3d) produce the correct result; and an example of a word whose etymological development was mispredicted by both distributions (Fig. 3b).

tative, frequently cited works can conflate with that of their literary successors. This paper introduces a latent variable model that captures such literary influences while simultaneously accounting for uncertainties in the temporal assignments. While the latter aspect is only of limited importance for Latin, the corpus language discussed in this paper, it is certainly relevant for many ancient corpora whose temporal structure is more disputed. Our discussion has shown that the model retrieves meaningful intellectual lineages and structures of word reuse (see Sec. 5.2) and performs on par with latent variable models specifically designed for capturing temporal topical trends (Sec. 5.3). In addition, the discussion of etymological derivations in Sec. 5.4 indicates that the linguistic distributions generated by the model are better able to describe certain aspects of language development than plain corpus distributions. Future extensions should incorporate a component that smoothes the temporal distributions [see e.g. 23], and they should consider non-temporal influence factors such as the geographic origin or genre of a text, as was proposed by Perrone et al. [49]. Given this outcome, we are planning to apply the mixture model on text traditions of ancient South Asia whose intellectual and diachronic structures are still not fully understood.

Acknowledgments

We thank Sabine Tittel for her help with digital resources for Romance languages and the three anonymous reviewers for their insightful comments. The authors were partly funded by the German Federal Ministry of Education and Research, FKZ 01UG2121.

References

- [1] J. E. Joseph, *Eloquence and Power: The Rise of Language Standards and Standard Languages*, Frances Pinter, London, 1987.
- [2] J. Clackson, *Classical Latin*, in: J. Clackson (Ed.), *A Companion to the Latin Language*, Blackwell Publishing, Maiden, MA, 2011, pp. 236–256.
- [3] J. Wackernagel, *Altindische Grammatik. I. Lautlehre*, Vandenhoeck und Ruprecht, Göttingen, 1896.
- [4] J. Lee, A computational model of text reuse in ancient literary texts, in: *Proceedings of the 45th ACL*, 2007, pp. 472–479.
- [5] M. J. Roberts, *The Hexameter Paraphrase in Late Antiquity: Origins and Applications to Biblical Texts*, University of Illinois, Urbana-Champaign, 1978.
- [6] G. Dinkova-Bruun, *Medieval Latin*, in: J. Clackson (Ed.), *A Companion to the Latin Language*, Blackwell Publishing, Maiden, MA, 2011, pp. 284–302.
- [7] J. Herman, *Vulgar Latin*, Pennsylvania State University Press, University Park, Pennsylvania, 2000.
- [8] E. Manjavacas, F. Karsdorp, M. Kestemont, A statistical foray into contextual aspects of intertextuality, in: F. Karsdorp, B. McGillivray, A. Nerghes, M. Wevers (Eds.), *Proceedings of the Workshop on Computational Humanities Research (CHR 2020)*, 2020, pp. 77–96.
- [9] S. Nehrlich, A method for the calculation of parallel passages for Buddhist Chinese sources based on million-scale Nearest Neighbor search, *Journal of the Japanese Association for Digital Humanities* 5 (2020) 132–153.
- [10] C. R. Stone, What plagiarism was not: Some preliminary observations on classical Chinese attitudes toward what the West calls intellectual property, *Marquette Law Review* 92 (2008) 199.
- [11] L. Dietz, S. Bickel, T. Scheffer, Unsupervised prediction of citation influences, in: *Proceedings of the 24th ICML*, 2007, pp. 233–240.
- [12] R. M. Nallapati, A. Ahmed, E. P. Xing, W. W. Cohen, Joint latent topic models for text and citations, in: *Proceedings of the 14th ACM SIGKDD*, 2008, pp. 542–550.
- [13] X. Wang, A. McCallum, X. Wei, Topical n-grams: Phrase and topic discovery, with an application to information retrieval, in: *Proceedings of the Seventh ICDM*, 2007, pp. 697–702.
- [14] D. M. Blei, J. D. Lafferty, Dynamic topic models, in: *Proceedings of the 23rd International Conference on Machine Learning*, 2006, pp. 113–120.
- [15] C. Wang, D. Blei, D. Heckerman, Continuous time dynamic topic models, in: *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, 2008, pp. 579–586.
- [16] B. Laurioux, *Cuisiner à l’antique: Apicius au Moyen Âge*, *Médiévales* (1994) 17–38.

- [17] P. Olivelle, The Early Upaniṣads. Annotated Text and Translation, Oxford University Press, Oxford, 1998.
- [18] O. Hellwig, Dating and stratifying a historical corpus with a Bayesian mixture model, in: Proceedings of LT4HALA, 2020, pp. 1–9.
- [19] D. Cohn, T. Hofmann, The missing link – A probabilistic model of document content and hypertext connectivity, in: Advances in Neural Information Processing Systems, 2001, pp. 430–436.
- [20] E. Erosheva, S. Fienberg, J. Lafferty, Mixed-membership models of scientific publications, Proceedings of the National Academy of Sciences 101 (2004) 5220–5227.
- [21] M. Rosen-Zvi, T. Griffiths, M. Steyvers, P. Smyth, The author-topic model for authors and documents, in: Proceedings of the 20th Conference on Uncertainty in AI, 2004, pp. 487–494.
- [22] E. Colledge, James of Voragine’s “Legenda Sancti Augustini” and its sources, Augustiniana 35 (1985) 281–314.
- [23] L. Frermann, M. Lapata, A Bayesian model of diachronic meaning change, Transactions of the Association for Computational Linguistics 4 (2016) 31–45.
- [24] X. Wang, A. McCallum, Topics over time: A non-Markov continuous-time model of topical trends, in: Proceedings of the 12th ACM SIGKDD International conference on Knowledge Discovery and Data Mining, 2006, pp. 424–433.
- [25] N. Kawamae, Trend analysis model: Trend consists of temporal words, topics, and timestamps, in: Proceedings of the fourth ACM International Conference on Web Search and Data Mining, 2011, pp. 317–326.
- [26] H. M. Wallach, Topic modeling: Beyond bag-of-words, in: Proceedings of the 23rd ICML, 2006, pp. 977–984.
- [27] M. Nokel, N. Loukachevitch, Accounting n-grams and multi-word terms can improve topic models, in: Proceedings of the 12th Workshop on Multiword Expressions, 2016, pp. 44–49.
- [28] T. L. Griffiths, M. Steyvers, J. B. Tenenbaum, Topics in semantic representation., Psychological Review 114 (2007) 211–244.
- [29] B. McGillivray, Methods in Latin Computational Linguistics, volume 1 of *Brill’s Studies in Historical Linguistics*, Brill, Leiden, 2014.
- [30] Y. Ouvrard, P. Verkerk, Collatinus & Eulexis: Latin & Greek dictionaries in the digital ages, in: Digital Classics III: Re-thinking Text Analysis, Center for Hellenic Studies/Harvard University, 2017.
- [31] B. Adamik, The periodization of Latin. An old question revisited, in: G. V. Haverling (Ed.), Latin Linguistics in the Early 21st Century, Uppsala Universitet, 2015, pp. 640–652.
- [32] T. L. Griffiths, M. Steyvers, Finding scientific topics, Proceedings of the National Academy of Sciences 101 (2004) 5228–5235.
- [33] D. Mimno, D. M. Blei, B. E. Engelhardt, Posterior predictive checks to quantify lack-of-fit in admixture models of latent population structure, Proceedings of the National Academy of Sciences 112 (2015) E3441–E3450.
- [34] G. Trompf, The concept of the Carolingian Renaissance, Journal of the History of Ideas 34 (1973) 3–26.
- [35] M. S. Kempshall, The virtues of rhetoric: Alcuin’s “Disputatio de rhetorica et de uirtutibus”, Anglo-Saxon England 37 (2008) 7–30.
- [36] F. Tutrone, Lucretius Franco-Hibernicus: Dicuil’s Liber de Astronomia and the Carolingian Reception of De Rerum Natura, Illinois Classical Studies 45 (2020) 224–252.

- [37] M. von Albrecht, *Geschichte der römischen Literatur*, volume 1, de Gruyter, Berlin, 2012.
- [38] H. Houghton, *The Latin New Testament*, Oxford University Press, Oxford, 2016.
- [39] D. Whitelock, *After Bede*, Bealls, Newcastle, 1978.
- [40] A. Kumar, M. Lease, J. Baldridge, Supervised language modeling for temporal resolution of texts, in: *Proceedings of the 20th ACM CIKM*, 2011, pp. 2069–2072.
- [41] L. Yao, D. Mimno, A. McCallum, Efficient methods for topic model inference on streaming document collections, in: *Proceedings of the 15th ACM SIGKDD*, 2009, pp. 937–946.
- [42] A. Stefenelli, Lexical stability, in: M. Maiden, J. C. Smith, A. Ledgeway (Eds.), *The Cambridge History of the Romance Languages. Volume I: Structures*, Cambridge University Press, Cambridge, 2011, pp. 564–584.
- [43] J. B. Solodow, *Latin Alive. The Survival of Latin in English and the Romance Languages*, Cambridge University Press, Cambridge, 2009.
- [44] T. Campbell, *Historical Linguistics*, Edinburgh University Press, Edinburgh, 2013.
- [45] N. Vincent, Continuity and change from Latin to Romance, in: J. Adams, N. Vincent (Eds.), *Early and Late Latin. Continuity or Change?*, Cambridge University Press, Cambridge, 2016, pp. 1–13.
- [46] M. Pagel, Q. D. Atkinson, A. Meade, Frequency of word-use predicts rates of lexical evolution throughout Indo-European history, *Nature* 449 (2007) 717–720.
- [47] W. Wu, D. Yarowsky, Computational etymology and word emergence, in: *Proceedings of The 12th Language Resources and Evaluation Conference*, 2020, pp. 3252–3259.
- [48] W. Meyer-Lübke, *Romanisches etymologisches Wörterbuch*, Winter, Heidelberg, 1935.
- [49] V. Perrone, M. Palma, S. Hengchen, A. Vatri, J. Q. Smith, B. McGillivray, GASC: Genre-aware semantic change for Ancient Greek, in: *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, 2019, pp. 56–66.