ISSN 2835-4138 (digital)
https://doi.org/10.5565/rev/isogloss.169

How fast did Cicero speak?
The speech rate of Classical Latin versus its Romance descendants

Daniel Stelzer
University of Illinois at Urbana-Champaign
stelzer3@illinois.edu

Received: 29-07-2021
Accepted: 29-08-2022
Published: 15-12-2022

How to cite Stelzer, Daniel. 2022. How fast did Cicero speak? The speech rate of Classical Latin versus its Romance descendants. RLLT18, eds. Jonathan MacDonald, Zsuzsanna Fagyal, Ander Beristain & Robin Turner. Special Issue of *Isogloss. Open Journal of Romance Linguistics* 8(4)/4, 1-24.
DOI: https://doi.org/10.5565/rev/isogloss.169

Abstract

While languages convey significantly different amounts of both information per syllable and syllables per second, recent research suggests that the product of these values—information conveyed per second—is much less variable. Using new methods of extrapolation and resampling, I was able to estimate the information conveyed per syllable in a written Classical Latin corpus. I was then able to use this cross-linguistic consistency to estimate the natural speech rate of Classical Latin, a language that has not been natively spoken for thousands of years. My analysis suggests that it was spoken at a rate significantly slower than modern Romance languages, fairly similar to modern English; a high-level consideration of historical sound changes in Romance supports this conclusion, lending additional credence to my results.

Keywords: Information density, speech rate, Latin, information theory, resampling.

1. Background

The idea that all languages are equally expressive—that, despite all the variation in phonology, syntax, and other aspects, every language is equally suited to communication and equally complex or sophisticated in its grammar—is ubiquitous in modern linguistics. Joseph and Newmeyer (2012) attribute the earliest expression of this concept to Humboldt in the 1820s, who claimed that "all [languages] contain all that is rigorously needed not only for the correctness, but the perfection of expression"[1]. By the end of the 19th century, this idea had gained support from the study of language evolution. Passy (1890: 227) suggested that language change is an eternal struggle between the tendencies "to get rid of what is superfluous" and "to highlight what is necessary"[2], preventing any overall increase or decrease of complexity.

Despite significant pushback, especially from those who bristled at the idea of equating European languages with those of Africa and the Americas, this idea slowly gained traction among the linguistic community. By 1955 it had made its way into the *Encyclopædia Britannica* article on "Language", now with a specific mention of 'complexity': "All languages of today are equally complex and equally adequate to express all the facets of the speakers' culture, and all can be expanded and modified as needed" (Trager 1955: 698). While the idea of 'primitive' versus 'sophisticated' languages persists in pop culture, this hypothesis of equal complexity is now put forth as a basic axiom in introductory textbooks. Akmajian et al. (2001: 8), for example, state plainly that "all known languages are at a similar level of complexity," and O'Grady et al. (2010: 8) assert that "linguists don't even try to rate languages as […] simple or complex."[3]

However, what exactly this universal 'expressiveness' (or 'complexity' or 'sophistication') *means* is far from obvious. Many typological studies, such as Maddieson (2005) and Shosted (2006), have tried to quantify the complexity of different aspects of grammar, looking for correlations between them (e.g., complicated phonology correlating with simpler morphology)—but results have generally been inconclusive.

[1] Quoted in Rémusat (1824: 8); translation from Joseph and Newmeyer (2012: 344).

[2] Translation from Joseph and Newmeyer (2012: 352).

[3] Outside of introductory textbooks, some linguists are more skeptical. Shosted (2006: 2) refers to it as "a claim that has been, until fairly recently, more a matter of dogma than of science", and Maddieson (2005: 216) suggests that "[s]uch a view seems to be based on the humanistic principle that all languages are to be held in equal esteem and are equally capable of serving the communicative demands placed on them. In rejecting the notion of 'primitive' languages linguists seem to infer that a principle of equal complexity must apply." On the flipside, other linguists support the claim for purely theoretical reasons: Chomsky (2004: 165-166) suggests that all languages "ought to" have the same overall budget for markedness, no matter how they spend it. Pellegrino, Coupé, and Marsico (2011: 540) suggest simply that "the assumption of an 'equal overall complexity' is ill-defined." For more discussion, see Joseph and Newmeyer (2012).

In the 1950s, this trend of research was given a new suite of tools from the burgeoning discipline of information theory. Shannon and Weaver (1949) introduced new mathematical models of information transmission based on concepts like entropy, noise, and channel capacity, and mathematically-inclined linguists soon started applying these models to spoken language. Karlgren (1962: 674,676), for example, suggested that "seemingly careless pronunciation" was actually "an efficient coding to fit the channel" of vocal transmission, and in particular that "there is an equilibrium between [Shannon's] information value on the one hand and duration and similar qualities of the realization on the other."

While Karlgren failed to find significant correlations between the lengths of words and the "carelessness" of speech (Karlgren 1962), the appeal of channel-focused models persisted. Lieberman (1963) attempted to measure the redundancy of various elements of a sentence, and found that the less redundant (i.e. more informative) elements were pronounced louder and longer. The word "nine" in the phrase "a stitch in time saves nine", for example, is extremely redundant: after hearing the five previous words, you know exactly what's coming next. The word "nine" in "that will be nine dollars", on the other hand, will significantly impact a listener's understanding of the sentence—and thus, it is pronounced more distinctly. Further experiments in the following decades supported Lieberman's results[4], and Aylett and Turk (2004) expanded these into a general principle, which they termed the "Smooth Signal Redundancy Hypothesis" (Aylett and Turk 2004: 34).

According to this hypothesis, speakers aim to have a similar level of redundancy across different parts of an utterance. If a phrase has a low level of linguistic redundancy (i.e. it is difficult to predict from context), it will show a higher level of acoustic redundancy (i.e. it will be pronounced clearly and carefully), and vice versa. The experimental evidence was promising, and Jaeger (2010) extended it into what he termed the 'Uniform Information Density Hypothesis'. Speakers, Jaeger predicted, should structure their syntax to convey a consistent rate of information over time: sculpting their language to fit what Shannon and Weaver (1949) would call the 'channel capacity' (Jaeger 2010: 3).

Preliminary experiments with both production and comprehension have supported this hypothesis, suggesting that Shannon and Weaver's information theory might be, in fact, a good model of human language (Jaeger 2010: 29; Meister et al. 2021: 970-971). If so, this could offer a new way of measuring languages' complexity, using information-theoretic methods. As Pellegrino, Coupé, and Marsico (2011: 539) put it, "[l]anguage is actually a communicative system whose primary function is to transmit information. The unity of all languages is probably to be found in this function, regardless of the different linguistic strategies on which they rely."

Since then, many authors have continued to examine language 'complexity' through this lens (Coupé et al. 2019; Oh 2015; Pellegrino, Coupé, and Marsico 2011). In particular, while acquiring acoustic data is still fairly time-consuming, access to written corpora has grown dramatically in recent years. Corpora with hundreds of millions of tokens have become common, and web-based English corpora easily reach the billions. Oh (2015) takes advantage of this data, analyzing the information density in written corpora from 18 languages and proposing a number of different metrics to investigate the Uniform Information Density Hypothesis in greater depth.

[4]See Aylett and Turk (2004: 32) for further references on these studies.

Some linguists, conversely, have questioned whether information-theoretic entropy is an appropriate metric for measuring the information content of natural language. After all, Shannon and Weaver (1949: 8) themselves specifically note that "[t]he word *information*, in this theory, is used in a special sense that must not be confused with its ordinary usage. In particular, *information* must not be confused with meaning." Information theory defines 'information' as a specific mathematical quantity, based on signals sent through a channel. But is that necessarily the same as the meaning that humans try to convey through language?

Pellegrino, Coupé, and Marsico (2011) searched for a different cross-linguistic measure of semantic density, in order to analyze how *meaning* (rather than Shannon's *information*) is encoded for verbal communication. The metric they came up with was based on the ratio of syllables in parallel corpora—in other words, the number of syllables that experienced translators use to convey the meaning of a particular text in a language, compared to the number of syllables needed to convey that same text in a control language.

Using this measure, they found a striking negative correlation between the density of semantic meaning per syllable ('meaning density'[5]) and the number of syllables spoken per second ('speech rate') in different languages. In the end, they rejected the hypothesis that the amount of meaning conveyed per second was uniform between languages—with a note that "[t]he very small size of the sample (N=seven languages) strongly limits the reliability of the results" (Pellegrino, Coupé, and Marsico 2011: 550). They were also limited by relatively small corpora—20 texts of five sentences each—and were not able to control for the effects of individual translators' style[6].

These results by Oh (2015) and Pellegrino, Coupé, and Marsico (2011) were brought together by Coupé et al. (2019), who first showed that one of Oh's corpus-based information density metrics was a good (and easier-to-calculate) proxy for Pellegrino, Coupé, and Marsico's meaning density[7], then applied it to a wide variety of languages using larger written corpora. They found that the amount of information conveyed per second ('information rate') was extremely consistent across the languages surveyed: while not perfectly constant (speech rate varies significantly by speaker and circumstance, for example), this average information rate varies much less by language than speech rate or information density, and seems to generally stay within a particular narrow band. They suggest that this 'optimal range' is a result of "universal communicative pressures characterizing the human-specific communication niche" (Coupé et al. 2019: 6). In other words, information rate is a property of how humans use language to communicate, on a larger scale than any individual language: "social and neurocognitive pressures […] define an optimal range for [information rate], around which the complex adaptive system (consisting of each language and its speakers) hovers" (Coupé et al. 2019: 6).

[5]'Meaning density' is my own terminology, to avoid ambiguity; Pellegrino, Coupé, and Marsico (2011) and Oh (2015) call it 'information density' (ID) or occasionally 'semantic information density', while Coupé et al. (2019) call it 'syntagmatic density of information ratio' (SDIR).

[6]It should be noted, though, that their corpora were significantly larger than those used in previous studies. For more details, see Pellegrino, Coupé, and Marsico (2011: 545).

[7]However, the sample size in Pellegrino, Coupé, and Marsico's study was fairly small. The relationship between 'information' and 'meaning' certainly merits further study, especially now that parallel corpora have become more available.

The present study is a new application of Coupé et al.'s results. Speech rate is normally calculated through recordings of native speakers, which is impossible for a dead language like Classical Latin[8]. However, Oh's methods of calculating information density are based on *written* corpora, which do exist for a number of extinct languages. If we assume the optimal information rate is constant across time and culture, can we calculate the information density from a corpus, and thereby 'reverse engineer' the speech rate?

2. Methods

2.1. Entropy

The concept of entropy, as used in this study, was first proposed by Shannon and Weaver (1949) as a way of quantifying information content. Shannon's model of communication involves an 'information source' emitting a series of discrete 'signals', one after another; in my model, the information source is a speaker (or writer) of a language, and the signals are syllables.

The 'entropy' of an information source then measures how much information, on average, each new signal conveys—or, equivalently, how much is *not known* about a signal before it is seen[9]. The original formulation from Shannon and Weaver (1949: 50) is now known specifically as the *Shannon entropy*:

$$H = -\sum_{x} P(x) \log P(x) \tag{1}$$

Here x is a type of signal and $P(x)$ is the probability of that signal. Intuitively, this means that information sources with more balanced probabilities will have higher entropy (if some types of signals are much more common than others, it is easier to guess what is coming next), and information sources with more types of signals will have higher entropy (if there are more possibilities, it is harder to guess what's coming next).

For my purpose, though, Oh (2015) suggests a slightly different model. In actual speech, signals do not exist in a vacuum devoid of context—mathematically, they are not independent. Consider, for example, English letters as signals. Without any context, the entropy is fairly high, since there are quite a lot of common letters to choose from. But right after a Q, the next letter is almost certain to be a U; a reader can be very confident what letter is coming next, giving the source an extremely low entropy in this situation. To model this, Shannon and Weaver (1949: 52) also propose what is now called the 'conditional entropy':

$$H_c = -\sum_{x,c} P(c,x) \log \frac{P(c,x)}{P(c)} \tag{2}$$

[8]Oh also investigated the speech rate of bilinguals in Oh, Coupé, and Pellegrino (2013). The results were not conclusive but suggest that the speech rate of L2 speakers can vary significantly from L1 speakers, meaning classicists who become fluent later in life cannot reliably tell us the speech rate of native speakers in ancient Rome.

[9]Hence the name 'entropy'. In statistical mechanics, entropy is a measure of uncertainty.

Here c is some representation of the context. In Oh's model, specifically, the signals are syllables of a language, and the context is the preceding syllable within the same word (making it a syllable bigram model)[10]. This is the metric used by Coupé et al. (2019), who term it ID (Information Density)[11]:

$$ID = -\sum_{\substack{\text{syllable,}\\ \text{context}}} P(\text{context, syllable}) \log \frac{P(\text{context, syllable})}{P(\text{context})} \tag{3}$$

My implementation follows Oh's (2015: 39) method exactly, using frequencies in a large corpus to approximate signal probabilities. Notably, I only consider context within a single word, not between words. This limitation was imposed by the corpora Oh used, many of which only provide individual word frequencies. I continue with it both to ensure my results can be compared directly against Coupé et al.'s (2019), who hewed similarly closely to Oh's methods, and to avoid the question of which boundaries (phrase, clause, sentence, paragraph, book) context should be able to cross[12].

2.2. Representation

The input to the bigram model discussed in section 2.1 is a broad[13] phonemic representation, with syllable boundaries marked. Since my corpus consists of plain text, I need a way of converting it to this phonemic representation. I accomplish this in three steps: augmenting the original orthography with additional data, converting the augmented orthography to a phonemic representation, and breaking this representation into syllables.

Fortunately, Classical Latin orthography is very close to phonemic. The Latin alphabet was still being modified during the Classical era, and deliberately-archaic spellings were rare[14]. This means that, for the most part, Classical orthography is thought to accurately represent the way the language was spoken at that time (Allen 1978: 9). However, some phonemic distinctions remain unrepresented, such as vowel quantity and vowels versus semivowels:

(4) ALIVM /a.li.um/ 'another'

(5) ALIVM /aː.li.um/ 'garlic'

(6) VOLVIT /wo.lu.it/ 'she wanted'

(7) VOLVIT /wol.wit/ 'it rolls'

[10]See Oh (2015: 41) for a worked example, using a toy language for demonstration; see Pellegrino, Coupé, and Marsico (2011: 545) for a discussion of using syllables versus phonemes as units.

[11]As mentioned in section 1, comparison with Pellegrino, Coupé, and Marsico's (2011) meaning density suggests this is a good measure of semantic content, though this relationship deserves further study.

[12]See, however, the comments in section 5.

[13]Though since the entropy is calculated on the syllable level rather than the phoneme level, and context is taken into account, the narrowness of the transcription is less important than for Shannon entropy.

[14]Archaisms were rare, but not unheard of. The archaic forms *quom* 'when' and *com* 'with' both became *cum* in Classical times due to a sound change, for example, leading some authors to use archaizing spellings to distinguish them. While the homonymy muddies the waters somewhat, the numbers here are illuminating: *quom* is attested 780 times in the corpus, *com* 3, and *cum* 55,738.

To account for this, I first convert the corpus to an 'augmented' or 'annotated' orthography, which distinguishes *a* from *ā*, *u* from *v*, and so on. This orthography is often used in introductory textbooks and is given here in *italics*, as opposed to SMALLCAPS for the original orthography found in the corpus; by design, it unambiguously represents all native phonemic differences.

Manually annotating a text in this way is generally straightforward—metered poetry and etymology reveal the quantity of most vowels, for example—but is extremely time-consuming. As a result, most documents in the corpus have never been manually annotated. For my analysis, I rely on an automatic annotation system developed by Winge (2015)[15], the heart of which is a customized version of Crane's (1991) Morpheus database. Winge (2015: 27) reports an accuracy exceeding 98% on classical texts, which is sufficient for my purposes.

The conversion from augmented orthography to phonemic representation is very regular, and mostly consists of handling quirks of the writing system. The augmented orthography is unambiguous, but still sometimes uses one letter for a sequence of multiple phonemes (*x* /ks/) or vice versa (*qu* /kʷ/)[16]. Following Allen (1978), I generally assume that a distinction in writing indicates a distinction in pronunciation—for example, the fact that the Greek letters Y and Z were borrowed during this period to transcribe loans implies that educated speakers really did pronounce them differently from I and S, and they should be treated as distinct phonemes. Similarly, spelling variations in transparent compounds like *ad-sum~as-sum* 'I am here' are taken to indicate actual variation between analogical and expected pronunciations[17].

For the most part, my transcription is phonemic, rather than phonetic. Stress, for example, does not seem to have been contrastive in Classical Latin—it is entirely predictable based on the segments of a word—and therefore I do not include it in my representation (Allen 1978: 83; Inst. Or.: I.5.30). There are, however, two specific types of non-phonemic detail reflected in my transcription.

First, I include any phonetic detail that could impact syllabification. Classical Latin does not seem to have made a phonemic distinction between /j/ and /jj/, for example, but /j/ between vowels within a root seems to have acted as both a coda and an onset[18]. So I represent phonemic /j/ in this environment as geminate /jj/, to ensure it is represented in both syllables. Intervocalic /w/, on the other hand, only seems to have acted as an onset[19] (e.g. *avis* [a.wis] 'bird', Spanish *ave*), so I always represent it as /w/.

[15]The system is available at `https://github.com/Alatius/latin-macronizer`, with an online demonstration at `https://alatius.com/macronizer/`.

[16]See `data/latin/process.py` for the full details.

[17]Compare Allen (1978: 22): "It is in fact uncertain to what extent in educated speech the analogical spellings may also have been reflected in pronunciation." My choice to take these spellings at face value was in part motivated by ease of implementation.

[18]In poetic meter, a syllable before /j/ behaves as closed, and its reflexes in certain Romance languages are consistently geminate: *major* [maj.jor] 'greater' > Italian *maggiore*.

[19]The one exception, based on meter, seems to be unassimilated Greek names in poetry. The difficulty of identifying these names consistently, plus uncertainty in how /w/ codas may have contrasted with /Vw/ diphthongs in actual speech, led me to ignore this in transcription. These names are infrequent enough that this is unlikely to have a significant impact on the results.

The other exception is complete neutralization. This is often reflected in Classical orthography already: for example, the distinction between /k/ and /kʷ/ is neutralized before /u/ and /uː/, giving *se**qu**-or* 'I follow' but *se**c**-undus* 'following'. Sometimes, though, morphological spellings hide this neutralization: *equus* 'horse' is written with *qu* under the influence of forms like *equī* 'of the horse', even though grammarians like Velius Longus (De Orth.: 59.2-8) indicate a pronunciation [ekus]. Similarly, *urbs* 'city' is written with *b* under the influence of forms like *urbis* 'of the city', but was almost certainly pronounced with a voiceless stop [urps], as suggested by Quintilian (Inst. Or.: I.7.7). I transcribe these words as /ekus/ and /urps/.

Syllabification, finally, is a much more thoroughly-studied topic. Classical metered poetry treats closed and open syllables differently, so formulating rules of syllabification has long been of interest to poets and poetry scholars, and syllable codas also impacted certain sound changes in Romance. For this I used an algorithm from CLTK (Johnson et al. 2014–2021), specially modified to remove certain hyperforeignisms specified by ancient authors[20].

The syllabified, near-phonemic transcription can then be converted into unigram and bigram frequency lists, which form the input to equation 3.

2.3. Extrapolation

In previous studies, Coupé et al. (2019) and Oh (2015) mostly used corpora with tens or hundreds of millions of tokens[21]. However, the entire surviving corpus of Classical Latin literature contains fewer than seven million tokens (Packard Humanities Institute 1991)—hundreds of times smaller than English Wikipedia[22], and on the same order of magnitude as a single month of the New York Times[23]. The first iteration of my study was limited to an even smaller corpus, with less than two million tokens. And while it seems clear that a larger corpus results in a more accurate estimate, it is not clear *how* large of a corpus is necessary, or whether mine is sufficient.

[20]For example, *pt* does not seem to ever have been a valid onset in Latin; poetic syllabifications like *ru-ptus* 'broken' are in imitation of Greek.

[21]See Oh (2015: 30-31) for a list of corpora used. Notably, while Oh did use some corpora that were significantly smaller (such as Robert's Wolof corpus with 0.07 million tokens), this likely affected the accuracy of the entropy estimates. Robert's corpus does not seem to be available online, but is discussed in Robert (2017), among other places.

[22]English Wikipedia consists of over 3.9 billion tokens at the time of initial writing, according to the "Size of Wikipedia" page (Wikimedia Foundation 2021).

[23]This is a rough back-of-the-envelope calculation, based on reports of averaging 150 articles per weekday, 250 articles per Sunday, and 622 words per article (Menendez-Alarcon 2012; Meyer 2016). These numbers are far from rigorous, but serve as an intuitive point of reference for the size of the Classical corpus.

One classic solution to the problem of limited data is bootstrapping, a way of artificially enlarging a dataset (Efron 1892; Oh 2015). But as Oh (2015: 55-56) demonstrates, bootstrapping does not seem to be a good tool for estimating entropy. In a natural-language corpus, Zipf's law (and its various extensions) predicts that a significant number of types will only occur once in the corpus—and even more valid types will never appear at all (Davis 2018). In Latin, for example, the word *audīverās* 'you had heard' appears only once[24], and its relative *audīverint* 'they might have heard' is completely unattested, presumably by sheer accident. Bootstrapping a dataset like this will very often change a frequency of 1 to 0, but can never change a frequency of 0 to 1, skewing the distribution. This may be the cause of some instability noted by Oh (2015: 56).

Notably, however, the issues with bootstrapping arise from sampling with replacement. By sampling *without* replacement, we can ensure the distribution is preserved. The result is always smaller than the original corpus, but by functionally discarding tokens at random, we can quite reliably create a smaller corpus that maintains the proper distribution. Oh (2015: 57) uses this technique to demonstrate that estimated information density increases sharply with corpus size, then appears to converge (see figure 1). I replicated her results for English and German, randomly discarding from the corpora to create smaller sub-corpora and calculating the information density from them. I then attempted to fit a curve to the convergence. In particular, the hyperbola shown in equation 8 fit extremely well[25]; it relates the estimated entropy (y) to the corpus size (x) with four parameters tuned through least-squares fitting. The results are shown in figure 1.

(8) $$y = a - b(x - c)^{-d}$$

To test the extrapolation, I randomly sampled two million tokens from the English and German corpora (without replacement), then extrapolated from these smaller sub-corpora (each approximately the size of our original Latin corpus). The results are shown in figure 2. For English, extrapolating from the smaller sub-corpus gave an information density of 7.00 bits per syllable, compared to 6.98 calculated from the sub-corpus, 6.98 calculated from the full corpus[26], or 6.99 extrapolated from the full corpus. For German, extrapolating from the smaller sub-corpus gave an information density of 6.11 bits per syllable, compared to 6.08 calculated from the sub-corpus, 6.08 calculated from the full corpus, or 6.10 extrapolated from the full corpus. This implies that two million tokens is sufficient for a good estimate on its own, but also gives me confidence in my method of extrapolation.

[24] Terence's *Phormio*, line 573: "So why did you stay there for so long, I must ask, once *you had heard* the news?"

[25] This curve was found empirically to have the best least-squares fit out of several tested, including other hyperbolic curves as well as exponential $a(1 - \exp(-b(x - c)))$ and logarithmic $a \ln(b(x - c))$.

[26] This number differs slightly from Oh's (2015: 61) 7.09 due to differences in the corpus; the exact corpus used by Oh was not available.

Figure 1. The convergence of estimated information density for German, approximated by equation 8.

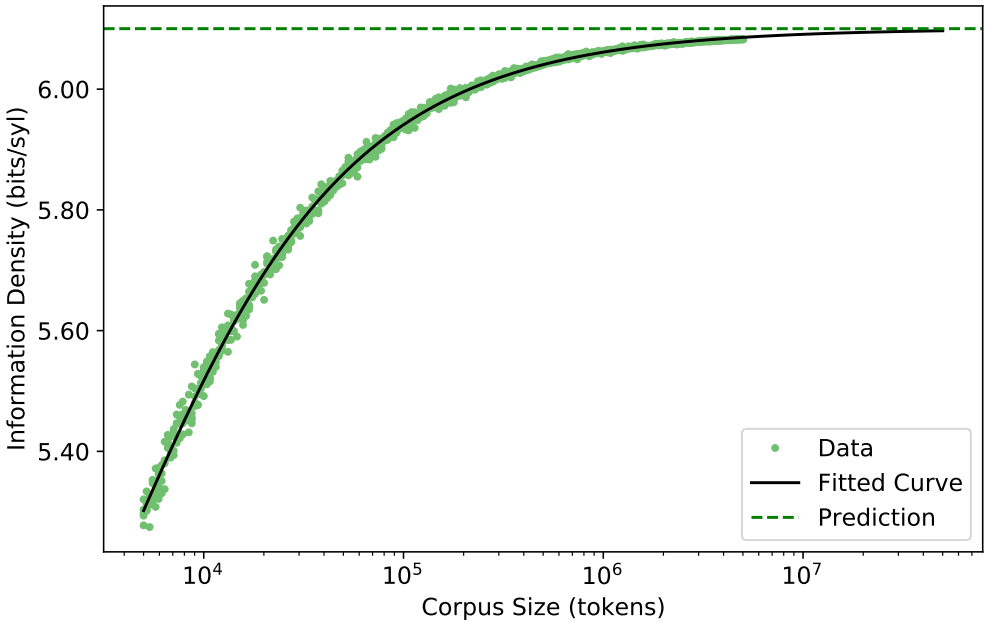
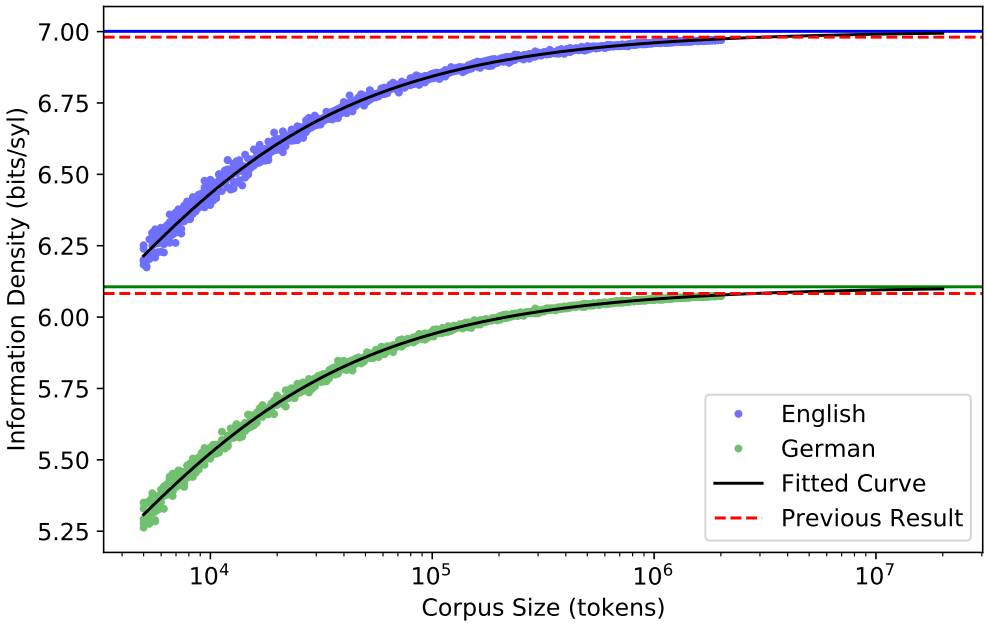


Figure 2. Information density extrapolated from small sub-corpora of English and German (solid line), compared to the result calculated from the entire corpus (dashed line). The difference suggests that expanding the full corpora by another order of magnitude would give us a slightly higher information density.



2.4. Jackknifing

Even if the number of tokens is sufficient for a good estimate, using the *entire* corpus of surviving literature raises another issue: how do we know that the corpus is representative of the language? When I discarded tokens from the English and German corpora, I discarded at random, ensuring the result followed the same distribution as the full corpus (and redoing the frequency calculations each time to account for the altered corpus size). But we have no way of knowing whether the works of literature that *didn't* survive had the same distribution as the ones that did[27].

In some ways, this is a problem that cannot be overcome. Barring a new archaeological find or rediscovered manuscript, the details of the literature that didn't survive are simply unknowable. But using the corpus I *do* have, I can analyze the effect of each individual source on the results, giving a sense of how much my result could be swayed by any particular lost source.

For this purpose, I assume that the differences between authors are more significant than the differences between works. In other words, the loss of Ovid's *Medea* is unlikely to make a significant impact on the calculations, since many other works by Ovid survive. The loss of the complete works of Cornelius Gallus, on the other hand, could matter a great deal, since he would have offered a completely different authorial voice and style—possibly with a much higher or lower information density than the others.

To estimate the impact of the limited pool of authors, I propose a new technique I term 'author jackknifing', named after jackknife resampling in statistics (Efron 1892). In this technique, I calculate the information density of the entire corpus, with one particular author removed—for example, I would calculate the information density of the entire corpus minus the works of Ovid, or the entire corpus minus the works of Livy. The distribution of these values, then, gives a sense of the impact an individual author could have, and I can take the standard deviation of this distribution as an approximation of the standard error. If this uncertainty is low, that suggests that a single author's style is unlikely to have a large impact on the results, and I can be more confident in my estimate.

3. Results

The corpus I used is the PHI Latin Corpus, published by the Packard Humanities Institute (1991). It contains, in their words, "essentially all Latin literary texts" from before 200 CE, plus a few later works that are deemed important and distinctly Classical in style[28]. Nearly every text that is recognizably Classical Latin and part of a published work is included, regardless of length or genre. In particular, I used version 5.3 as distributed on CD, in conjunction with CLTK's index (Johnson et al. 2014–2021), as it made it easier to access full texts than the newer web interface. This version of the corpus includes 329,228 types and 7,240,273 tokens[29], from 362 authors.

[27] There is also a question of how accurately written literature represents the spoken language, but this is an issue with any written corpus. Following Oh (2015) and Coupé et al. (2019), we ignore it here.

[28] For example, the corpus includes the commentaries of Servius Honoratus, from the fourth century CE.

[29] These numbers differ from the ones in table 1 because the values here include Justinian and table 1 does not.

However, most of these 362 authors have relatively little contribution—for example, Cornelius Dolabella's only surviving text is the two words *mortem ferre* 'to bring death'[30]. Including these would make author jackknifing somewhat useless, since it is clear that two words from an unknown author cannot significantly impact the entropy of a seven-million-word corpus. So, for the purposes of jackknifing specifically, I included only authors who contributed more than 100,000 tokens, as shown in table 1.

Table 1. The composition of the corpus

Author	Genre	Century	Types	Tokens
Total			321,447	6,387,500
Cicero	Everything	1st BCE	85,020	1,165,502
Justinian	*Law*	*6th CE*	*40,599*	*852,973*
Livy	History	1st BCE	55,308	520,674
Pliny	Science	1st CE	68,237	392,178
Servius	Grammar	4th CE	52,524	373,819
Seneca	Philosophy	1st CE	50,723	362,937
Quintilian	Rhetoric	1st CE	40,092	321,209
Ovid	Poetry	1st BCE	36,787	222,745
Plautus	Comedy	3rd BCE	27,352	166,390
Tacitus	History	1st CE	33,744	161,368
Gellius	Notes	2nd CE	24,636	118,021
Columella	Agriculture	1st CE	26,019	115,811
"S. H. A."[31]	Biographies	4th CE?	23,892	107,893
Apuleius	Novel	2nd CE	30,735	103,901
Celsus	Religion	2nd CE	15,736	102,035
Others			189,907	2,153,017

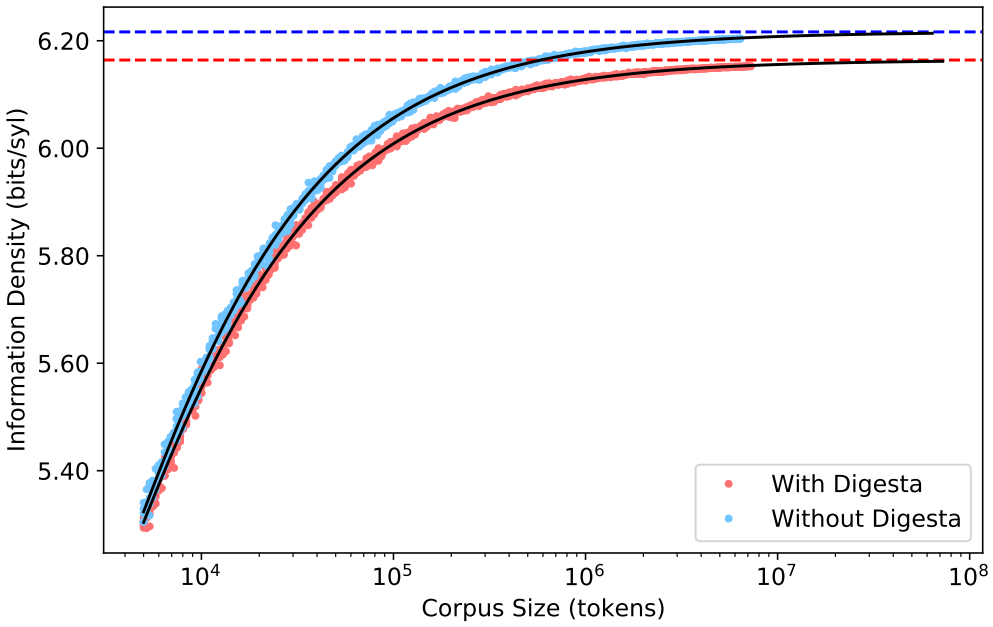
Using this corpus for analysis, and these fifteen authors for jackknifing, two outliers immediately became apparent.

The first involves Justinian. The *Digesta* of Justinian is one of the later (post-200-CE) works included in the CLTK corpus: a fifty-volume compilation of legal precedents and decisions. Since many of these precedents are from the Classical period, it makes some sense to include them in the corpus. However, they are also extremely formulaic and repetitive (note the very low type/token ratio in table 1)—to the point that they significantly lower the overall information density of the corpus, as shown in figure 3. Since the *Digesta* was compiled much later than the other works in the corpora, we feel comfortable excluding it as an outlier.

[30]Quoted in Quintilian (Inst. Or.: VIII.2.4). A fair number of authors are primarily (or only) known to us through ancient quotations, raising questions about their authenticity or their usefulness for corpus analysis; discussion of some of these problems, and the general practices of quotation in this era, can be found in Hoek (1996). I follow the decisions of the Packard Humanities Institute in this area, separating out quoted authors wherever the compilers of the corpus deemed it useful to do so.

[31]*Scrīptorēs Historiae Augustae*, literally the "authors of the Augustan History". The actual identity of the author, or authors, is unknown.

Figure 3. Estimated information density with and without the *Digesta*



The second involves Cicero. He was an incredibly prolific writer, and due to the praise of early Christian church fathers, his works were preserved better than most others'—it is estimated that over 75% of the surviving writings from his lifetime are his (Harrison 2008). Even looking at other time periods, his works make up over 18% of my corpus. As such, he has a much greater impact on my information density estimate than most other authors, through sheer volume (as can be seen in figure 5).

Arguments could be made to exclude his works from the corpus as an outlier, due to this impact, and also what some have called his "extremely poor vocabulary"[32] (Albrecht 2017: 136). Arguments could also be made to include his works in the total corpus but exclude them from the jackknifing, since some consider Cicero's works definitional to the 'Classical' style, rather than being just another author (Albrecht 2017: 136). In the end, the decision was made to include his works both in the corpus *and* in the jackknifing. Even if his vocabulary was deliberately limited, his works are still an important example of Latin written in the Classical period, and from a descriptive standpoint I believe his authorial style should factor into the uncertainty the same as any other author's.

Using this corpus (with Justinian excluded but Cicero included), the results of the extrapolation can be seen in figure 4. The information density calculated from the entire corpus is 6.204 bits per syllable, extrapolated to 6.216 bits per syllable.

The results of author jackknifing are shown in figure 5. The mean of the extrapolated entropy of the fourteen resampled corpora is 6.220 bits per syllable, with a standard deviation of 0.0198. This standard deviation approximates how much any particular lost author could impact the results, and I take it as a measure of the uncertainty in the estimate.

[32]This is not to suggest that he didn't know many words, but rather that he deliberately strove for consistency and clarity in his vocabulary, discarding many synonyms and alternate constructions as a result. In his works on oratory, he describes this as a key component of good speech.

Figure 4. Estimated information density from the entire corpus

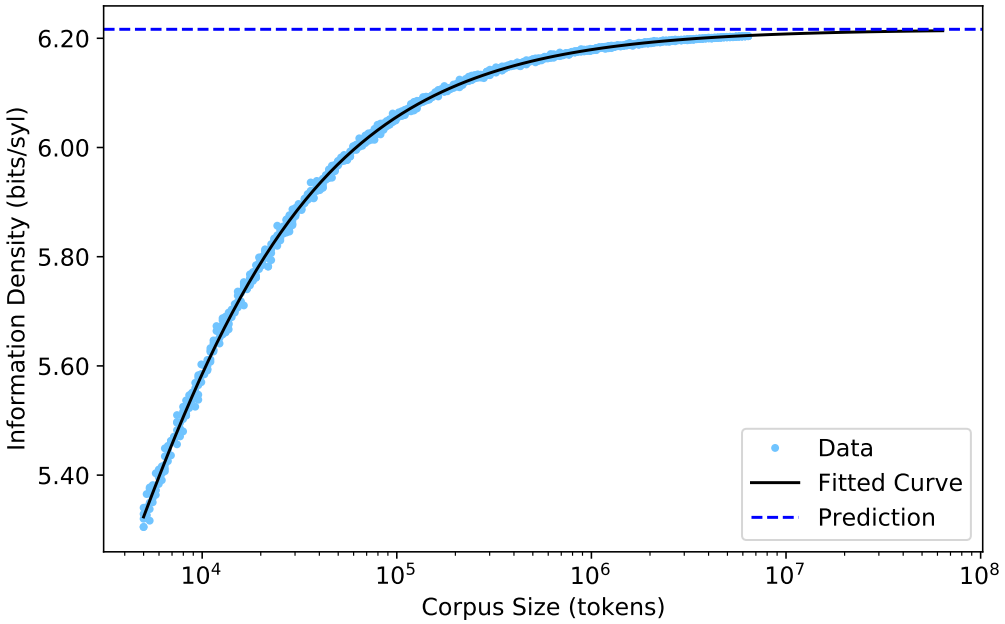
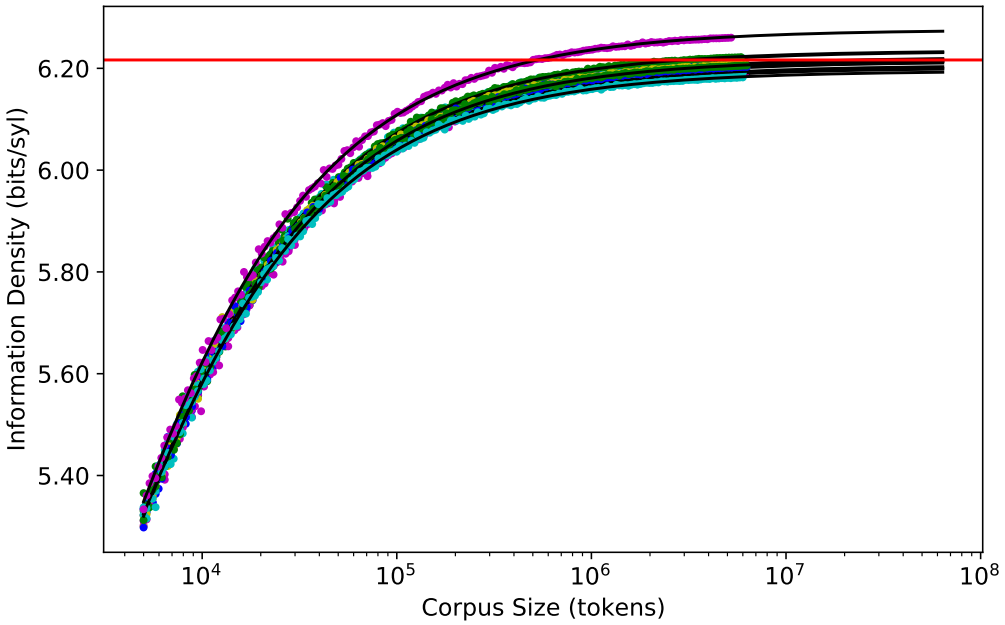


Figure 5. Results of author jackknifing. The visible outlier (in purple) is the corpus without Cicero.



According to Coupé et al. (2019), the mean information rate across languages is 39.15 bits per second, with a standard deviation of 5.10. From this, I calculate a reconstructed speech rate:

$$SR = \frac{IR}{ID} = 6.29 \text{ syl/sec} \tag{9}$$

$$\sigma_{SR} = \sqrt{\left(\frac{\sigma_{IR}}{ID}\right)^2 + \left(\frac{\sigma_{ID}}{ID}\right)^2} = 0.82 \tag{10}$$

Notably, the uncertainty in the information density value (σ_{ID}) is so small compared to the natural variation in IR (σ_{IR}) that it becomes negligible. The effect of different authorial styles on the information density estimate is completely drowned out by the size of the 'optimal range' for IR, suggesting that this corpus is diverse enough to be confident in the results.

4. Discussion

Compared to Coupé et al.'s (2019) empirical data, these results—6.29 syllables per second with a standard deviation of 0.82—seem quite reasonable. The rate is fairly similar to that of English, with some other languages being much faster and others much slower. The uncertainty is significantly smaller than many of the differences between languages, showing that our results are precise enough to be meaningful. And the uncertainty is also fairly similar to the variations reported in Coupé et al.'s experiment.

Figure 6 compares our reconstructed speech rate for Classical Latin against the data gathered by Coupé et al. (2019) for Vietnamese, English, and Japanese. Tickmarks indicate the mean and one standard deviation above and below; for living languages, the 'violin' of the plot shows the distribution of measurements of actual speech rate, while for Latin, it's approximated by a normal distribution with our calculated mean and standard deviation.

4.1. Comparison

Now that we know the estimated speech rate is plausible for a language in general, a more interesting question arises: how does it compare to its descendants, the modern Romance languages? Figure 7 compares my reconstructed values for Latin against the measured values for the four Romance languages included in Coupé et al.'s (2019) study: Catalan, French, Italian, and Spanish.

Notably, Latin seems to be spoken significantly slower than all the Romance languages tested, often by at least one standard deviation. Spanish, the fastest of them, is more than an entire syllable per second faster.

Figure 6. Our reconstructed speech rate for Classical Latin, compared to measured speech rates of living languages: Vietnamese, English, and Japanese

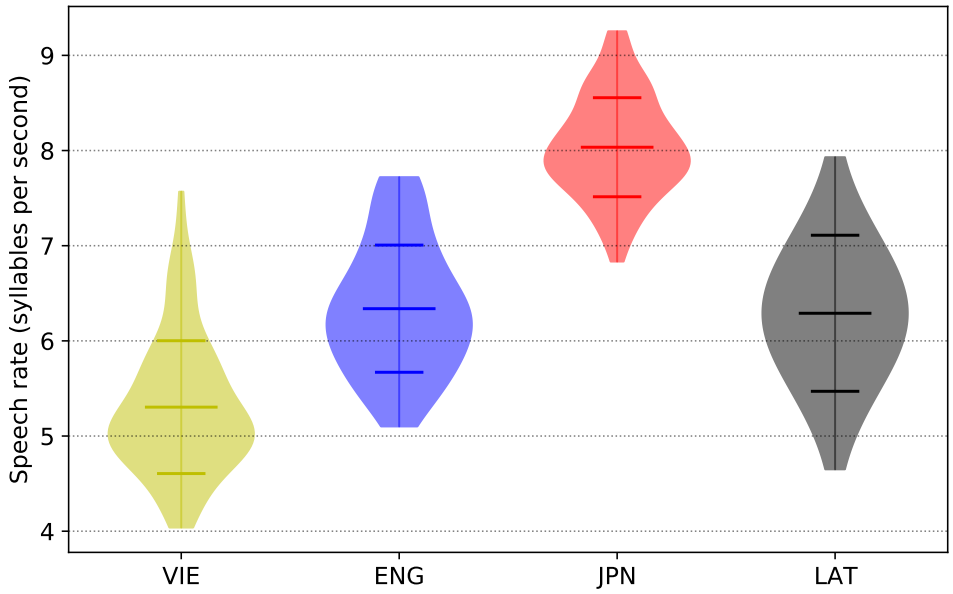
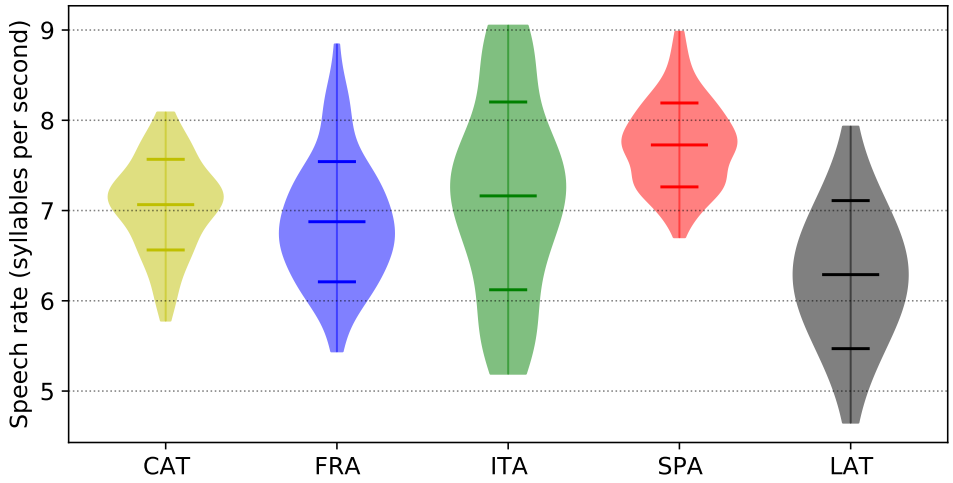


Figure 7. The speech rate of Classical Latin compared to modern Romance languages: Catalan, French, Italian, and Spanish



The next question is, are there phonological reasons for this? In other words, are there significant phonological differences between Classical Latin and its modern descendants that would explain this significant difference in speech rate? In table 2 I reproduce some phonological data from Oh (2015) for comparison[33]. 'C', 'V', and 'T' indicate the size of the inventory of consonants, vowels, and tone/stress features; 'Complexity' and 'Index' are measures of syllable complexity proposed by Maddieson (2013) and Maddieson et al. (2013) respectively (the 'index' being roughly equivalent to the maximum number of segments in a single syllable); and 'SE' and 'ID' are the Shannon entropy and information density reported by Oh (2015), both in bits per syllable.

Table 2. Phonological statistics for Classical Latin compared to modern Romance languages: number of consonants, vowels, and tone/stress features; complexity and complexity index; Shannon entropy and information density.

Language	C	V	T	Complexity	Index	SE	ID
Latin	21	17	0	Complex	7	8.71	6.22
Catalan	25	8	2	Moderate	4	8.10	5.49
French	22	15	0	Complex	7	8.39	6.68
Italian	27	7	1	Complex	6	8.32	5.29
Spanish	27	5	1	Moderate	5	8.32	5.43

Romance data taken from Oh (2015: 44-45).

Notably, the Shannon entropy (SE) of Classical Latin—that is, the syllable entropy without context, as formulated in equation 1—is much closer to that of modern Romance languages than the information density (ID). I attribute this mainly to the presence of stress. Phonemic stress, as found in many Romance languages, increases the number of possible syllables immensely (since, for example, stressed *á* and unstressed *a* become phonologically distinct syllables). This increase in the syllable inventory then greatly increases the Shannon entropy. But the information density incorporates context as well, and in context, stress matters much less—in stress-accent languages, there tends to be one and only one ictus per word, not an independent binary "stressed/unstressed" property for each syllable. Latin lacks phonemic stress but has phonemic vowel length, which is a property of each individual vowel rather than the word: *pila* 'ball', *pīla* 'mortar', *pilā* 'with a ball', *pīlā* 'with a mortar'. So while the effects of stress and vowel length on the Shannon entropy are similar, their effects on the information density are very different.

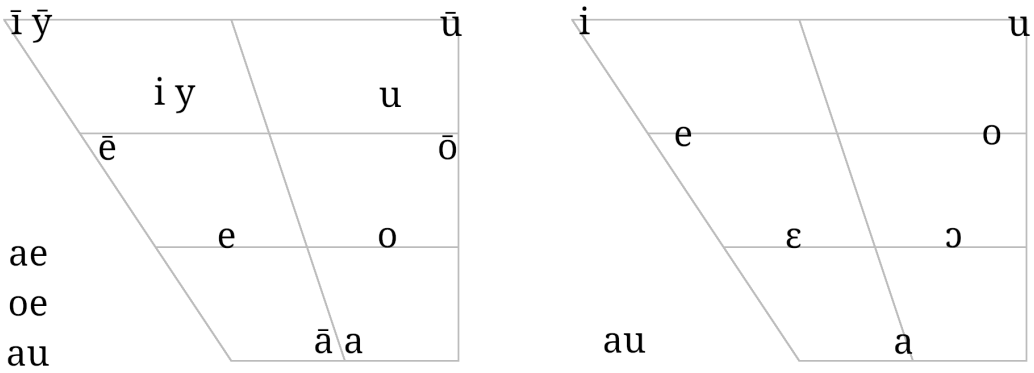
[33] An estimate of syllable inventory size based on corpus measurements is also reported by Oh (2015), but not included here. Her method of estimation involves looking at the most frequent lemmata in the corpus, and our Latin corpus has not been lemmatized.

The 'Complexity' and 'Index' values, while fairly similar here, can also obscure significant differences in syllable structure. Latin generally allows up to three consonants in a coda: *urbs* /urps/ 'city', *calx* /kalks/ 'chalk'. Italian, on the other hand, does not allow any clusters in codas (Hall 1944). The reason for the similar 'Index' and 'Complexity' values is the analysis of diphthongs—the /ae̯/ in Latin *saepe* 'often' is taken as a single segment, while the /ai̯/ in Italian *assai* 'very' is taken as two segments, for phonological reasons—and the number of possible diphthongs is much smaller than the number of possible consonant clusters.

4.2. Historical Phonology

Leaving aside Maddieson et al.'s (2013) index, I suggest a broader look at the history of the Romance languages, and what sorts of common phonological changes these languages shared.

Figure 8. The vowel inventories of Classical Latin (left) and early Romance (right)



Vowel changes: Classical Latin had twelve phonemic monophthongs, /i y e a o u/ with a binary length distinction, and at least[34] three phonemic diphthongs, /ae̯ oe̯ au̯/. Early in Proto-Romance, the length distinction became a quality distinction, and a series of mergers resulted in seven phonemic monophthongs /i e ɛ a ɔ o u/ and one diphthong /au̯/, as shown in figure 8. Alkire and Rosen (2010) refer to this as the 'Great Merger', and its effects can be seen in almost all Romance languages[35]; most underwent further mergers, especially when unstressed (Alkire and Rosen 2010; Boyd-Bowman 1980). The main result of this was to cut the number of possible syllables nearly in half.

Cluster breaking: While Classical Latin allowed /s/ before stops in onsets, Western Romance varieties historically did not.

(11) *strictus* 'tight' → Spanish *estrecho*

(12) *sponsa* 'fiancée' → Old French *espose* → French *épouse*

(13) *scriptus* 'written' → Catalan *escrit*

[34]Some analyses, such as Allen (1978), consider other diphthongs phonemic as well, but others, such as Alkire and Rosen (2010) and Boyd-Bowman (1980), include only these three. In this paper, I follow Allen for most matters of Classical phonology and phonetics (including table 2), but Alkire and Rosen (2010) specifically in discussions of historical development (including figure 8).

[35]The famous exception is Sardinian.

This prohibition eventually weakened in some languages (such as Modern French), and the epenthetic vowels mostly disappeared in Italian—*stretto*, *sposa*, *scritto*—though relics like *iscritto* survive in fossilized phrases[36] (Alkire and Rosen 2010). But the primary effect was to break up inherited consonant clusters, increasing the number of syllables per word and decreasing the number of possible onsets.

Coda loss: Classical Latin allowed a wide variety of coda consonants, including stops, nasals, and fricatives. Many of these disappeared in the development of Romance, especially word-finally:

(14) *habētis* 'you all have' → Italian *avete*

(15) *ferrum* 'iron' → Catalan *ferro*

(16) *amābat* 'she was loving' → Spanish *amaba*

Certain codas were already being lost in the Classical period—poetry and inscriptions indicate that coda /m/, while phonologically present, was no longer realized as a separate segment[37]—and some Romance branches took this process farther than others. Latin *servōs* 'slaves', for instance, became Spanish *siervos* but Italian *servi* (via an intermediate **servoj*). This contributed, again, to a decrease in the inventory of possible syllables.

Consonant loss and gain: Some Classical consonants merged or disappeared without a trace in Romance, while some others arose out of previously non-phonemic contrasts:

(17) *hortus* 'garden' → Old French *ort*

(18) *chorus* 'chorus' → Spanish *coro*

(19) *fāgeus* /faːgeus/ 'beech' → Italian *faggio* /faddʒo/

The result is a similar number of consonants in Classical Latin and its descendants, as reported in table 2. While the inventory of consonants varies significantly, its size is similar between Latin and the Romance languages surveyed.

Syncope: Many unstressed medial vowels were deleted very early in the history of Romance. This change happened early enough to appear in inscriptions, and in a few cases the effects are even visible during the Classical period:

(20) *calida* 'hot' → Latin *calda*[38] → Italian *calda*

(21) *viridis* 'green' → Spanish *verde*

(22) *altera* 'other' → Old French *altre* → French *autre*

This created more consonant clusters—but, crucially, did not significantly increase the inventory of syllables, or allow syllables to appear in more contexts. All of these clusters and syllables existed in the language before the syncope process took place.

[36]Some older speakers have also been reported to preserve the epenthetic vowel when the word follows a consonant: *alla Svizzera* but *in Isvizzera*.

[37]In other words, it was probably realized as nasalization of the preceding vowel instead of as a consonant [m].

[38]The emperor Augustus is quoted by Quintilian (Inst. Or.: I.6.19) as calling the longer form "superfluous" and telling his grandson to avoid it.

French vowel changes: French is one of the less conservative Romance languages, phonologically, and in particular it has a significantly larger vowel inventory than Italian, Spanish, or Catalan. These vowels stem from a variety of dramatic changes from early Gallo-Romance, often conditioned by surrounding consonants which later disappeared (Pope 1934; Boyd-Bowman 1980):

(23) *fo**re**stem* 'forest' → *for**ê**t* /fɔʁɛ/

(24) ***den**tem* 'tooth' → ***den**t* /dɑ̃/

(25) *pedem* 'foot' → *pied* /pje/

The result is a much larger vowel inventory than the other Romance languages surveyed. This increased the French syllable inventory significantly; I hypothesize that this contributed to the language's relatively high information density, as shown in table 2.

In summary, I believe there are several historical reasons to expect modern Romance languages to have smaller syllable inventories than Latin. This reduces the informational load on each syllable and the difficulty of recognizing syllables, allowing the language to be spoken faster. While syllable frequency likely plays an important role as well, it is notable that the Romance language closest in speech rate to Latin, out of those surveyed, is French—which has a significantly larger vowel inventory than the rest.

5. Conclusion

In this study, I used a variety of tools to analyze the information density of Classical Latin from a written corpus. Applying my new methods of extrapolation and author jackknifing, I came up with an information density of 6.216 bits per syllable, with a standard error of 0.198. Applying these values to Coupé et al.'s (2019) information rate distribution, I predict a mean speech rate of 6.29 syllables per second, with a standard deviation of 0.82—significantly slower than the modern Romance languages tested by Coupé et al. I believe these results make sense, based on a broad overview of historical developments in Romance.

I believe these methods of extrapolation can be applied to other corpora. While I demonstrated that two million tokens is enough, more research is required to determine how small a corpus is sufficient for a good extrapolation. Future work might put a lower bound on this, and potentially apply it to other languages with less written data available.

I have also looked at a handful of prominent historical sound changes to explain these results, but in a language family as well-understood as Romance, this has barely scratched the surface of the potential. With a more detailed investigation of one particular language's historical phonology, it might be possible to trace the evolution of the language's speech rate over time and quantify the effect of particular phonological changes.

Finally, the Latin corpus may offer an opportunity to expand on Oh (2015) and Coupé et al.'s (2019) work. They relied solely on within-word context due to the limitations of the available corpora—for many languages, it can be difficult to find good corpora that are both phonemically annotated and include broader context. But Latin orthography is quite close to phonemic, as discussed in section 2.2, and between-word sandhi effects in Latin have long been studied for the same reason as syllabification: they have a significant impact on metered poetry. This could make it a good test subject for the effects of inter-word context.

The recent advances in the study of linguistic information are exciting, and I believe they have significant potential. Reconstructing phonetic properties of long-dead languages may be only the beginning.

Supplemental Materials

The source code used for this analysis can be found at https://github.com/dstelzer/latin-speech-rate.

Acknowledgements

I would like to thank Ryan Shosted for advice and guidance on every part of this project, Yoon Mi Oh for pointers on the entropy calculations, and Tim Stelzer and Ada Stelzer for mathematical assistance.

I would also like to thank the organizers and reviewers of the LSRL 51 conference where this work was first presented; Keith Tse, Benjamin Tucker, William Balla-Johnson, José Ignacio Hualde, and Ed Rubin for their questions and comments at that conference; and three anonymous reviewers for their feedback and insights.

References

Akmajian, Adrian, Richard A. Demers, Ann K. Farmer, and Robert M. Harnish. 2001. *Linguistics: An Introduction to Language and Communication*. 5th ed. MIT Press.

Albrecht, Michael von. 2017. *Cicero's Style: A Synopsis. Followed by Selected Analytic Studies*. Leiden, The Netherlands: Brill. ISBN: 978-90-47-40197-1. DOI: 10.1163/9789047401971. URL: https://brill.com/view/title/8198.

Alkire, Ti and Carol Rosen. 2010. *Romance Languages: A Historical Introduction*. Cambridge University Press.

Allen, W. Sidney. 1978. *Vox Latina: A Guide to the Pronunciation of Classical Latin*. 2nd ed. Cambridge University Press. DOI: 10.1017/CBO9780511620348.

Aylett, Matthew and Alice Turk. 2004. "The Smooth Signal Redundancy Hypothesis. A Functional Explanation for Relationships between Redundancy, Prosodic Prominence, and Duration in Spontaneous Speech". In: *Language and Speech* 47.1. PMID: 15298329: 31–56. DOI: 10.1177/00238309040470010201.

Boyd-Bowman, Peter. 1980. *From Latin to Romance in Sound Charts*. Georgetown University Press. ISBN: 9780878400775. URL: https://books.google.com/books?id=z0RbXWRbLHIC.

Chomsky, Noam. 2004. *The Generative Enterprise Revisited: Discussions with Riny Huybregts, Henk van Riemsdijk, Naoki Fukui and Mihoko Zushi*. Berlin: Mouton de Gruyter.

Coupé, Christophe, Yoon Mi Oh, Dan Dediu, and François Pellegrino. 2019. "Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche". In: *Science Advances* 5.9. DOI: 10.1126/sciadv.aaw2594. eprint: https://advances.sciencemag.org/content/5/9/eaaw2594.full.pdf. URL: https://advances.sciencemag.org/content/5/9/eaaw2594.

Crane, Gregory. Jan. 1991. "Generating and Parsing Classical Greek". In: *Literary and Linguistic Computing* 6.4: 243–245. ISSN: 0268-1145. DOI: 10.1093/llc/6.4.243. eprint: https://academic.oup.com/dsh/article-pdf/6/4/243/10889452/243.pdf.

Davis, Victor. 2018. "Types, Tokens, and Hapaxes: A New Heap's Law". In: *Glottotheory* 9.2: 113–129. DOI: doi:10.1515/glot-2018-0014. eprint: https://arxiv.org/abs/1901.00521.

Efron, Bradley. 1892. *The Jackknife, the Bootstrap and Other Resampling Plans*. DOI: 10.1137/1.9781611970319. URL: https://epubs.siam.org/doi/book/10.1137/1.9781611970319.

Hall, Robert A. 1944. "Italian Phonemes and Orthography". In: *Italica* 21.2: 72–82. ISSN: 00213020. URL: http://www.jstor.org/stable/475860.

Harrison, S. 2008. *A Companion to Latin Literature*. Blackwell Companions to the Ancient World. Wiley. ISBN: 9781405137379.

Hoek, Annewies Van Den. 1996. "Techniques of Quotation in Clement of Alexandria. A View of Ancient Literary Working Methods". In: *Vigiliae Christianae* 50.3: 223–243. ISSN: 00426032. URL: http://www.jstor.org/stable/1584076.

Jaeger, T. Florian. Aug. 2010. "Redundancy and reduction. Speakers manage syntactic information density". In: *Cogn Psychol* 61.1: 23–62. DOI: 10.1016/j.cogpsych.2010.02.002. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2896231/.

Johnson, Kyle P., Patrick Burns, John Stewart, and Todd Cook. 2014–2021. *CLTK: The Classical Language Toolkit*. URL: https://github.com/cltk/cltk.

Joseph, John E. and Frederick J. Newmeyer. 2012. "'All Languages Are Equally Complex': The rise and fall of a consensus". In: *Historiographia Linguistica* 39.2-3: 341–368. ISSN: 0302-5160. DOI: 10.1075/hl.39.2-3.08jos. URL: https://www.jbe-platform.com/content/journals/10.1075/hl.39.2-3.08jos.

Karlgren, Hans. 1962. "Speech Rate and Information Theory". In: *Proceedings of the Fourth International Congress of Phonetic Sciences*.

Lieberman, Philip. 1963. "Some Effects of Semantic and Grammatical Context on the Production and Perception of Speech". In: *Language and Speech* 6.3: 172–187. DOI: 10.1177/002383096300600306.

Longus, Velius. c. 150. *De Orthographia*. URL: https://latin.packhum.org/loc/1374/1/0.

Maddieson, Ian. 2005. "Correlating Phonological Complexity: Data and Validation". In:

DOI: 10.5070/P795m171v6. URL: https://escholarship.org/uc/item/95m171v6.

— 2013. "Syllable Structure". In: *The World Atlas of Language Structures Online*. Ed. by Matthew S. Dryer and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology. URL: https://wals.info/chapter/12.

Maddieson, Ian, S. Flavier, E. Marsico, C. Coupé, and F. Pellegrino. 2013. "LAPSyD: Lyon-Albuquerque Phonological Systems Database". In: *Proceedings of Interspeech 2013*: 3022–3026.

Meister, Clara, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. Nov. 2021. "Revisiting the Uniform Information Density Hypothesis". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics: 963–980. DOI: 10.18653/v1/2021.emnlp-main.74. eprint: https://arxiv.org/abs/2109.1163. URL: https://aclanthology.org/2021.emnlp-main.74.

Menendez-Alarcon, Antonio. Jan. 2012. "Newspapers Coverage of Spain and the United States: A Comparative Analysis". In: *Sociology Mind* 2: 67–74. DOI: 10.4236/sm.2012.21009. URL: https://www.researchgate.net/publication/228446950_Newspapers_Coverage_of_Spain_and_the_United_States_A_Comparative_Analysis.

Meyer, Robinson. May 2016. "How Many Stories Do Newspapers Publish Per Day? A look at how The New York Times, The Wall Street Journal, the Washington Post, and BuzzFeed compare". In: *The Atlantic*. URL: https://www.theatlantic.com/technology/archive/2016/05/how-many-stories-do-newspapers-publish-per-day/483845/.

O'Grady, William, John Archibald, Mark Aronoff, and Janie Rees-Miller. 2010. *Contemporary Linguistics: An Introduction*. 6th ed. Bedford/St. Martin's.

Oh, Yoon Mi. Oct. 2015. "Linguistic Complexity and Information: Quantitative Approaches". PhD thesis. University of Lyon. URL: http://www.ddl.cnrs.fr/fulltext/Yoonmi/Oh_2015_1.pdf.

Oh, Yoon Mi, Christophe Coupé, and François Pellegrino. 2013. "Effect of Bilingualism on Speech Rate. The Case of Catalan and Basque Bilinguals in Spain". In: *2013 International Congress of Linguists (ICL)*. URL: http://hdl.handle.net/10722/283450.

Packard Humanities Institute. 1991. *PHI Latin Texts*. URL: https://latin.packhum.org/.

Passy, Paul. 1890. *Étude sur les changements phonétiques et leurs caractères généraux*. Paris: Firmin-Didot.

Pellegrino, François, Christophe Coupé, and Egidio Marsico. 2011. "A Cross-Language Perspective on Speech Information Rate". In: *Language* 87.3: 539–558. ISSN: 00978507, 15350665. URL: http://www.jstor.org/stable/23011654.

Pope, Mildred Katharine. 1934. *From Latin to Modern French with Especial Consideration of Anglo-Norman: Phonology and Morphology*. French series no.6. Manchester University Press. ISBN: 9780719001765. URL: https://books.google.com/books?id=K9JRAQAAIAAJ.

Quintilianus, Marcus Fabius. c. 95. *Institutio Oratoria*. URL: https://latin.packhum.org/loc/1002/1/0.

Rémusat, Jean Pierre Abel. 1824. "Review of Humboldt (1825, 1823-1824)". In: *Journal Asiatique* (5): 51–61.

Robert, Stéphane. 2017. *The Wolof Corpus and functional database on predication of the Cortypo project (Constitution de Corpus Oraux pour des recherches Typologiques / Designing spoken corpora for cross-linguistic research)*. URL: https://hal.archives-ouvertes.fr/hal-01678100.

Shannon, Claude E. and Warren Weaver. 1949. *The Mathematical Theory of Communication*. Illini books. University of Illinois Press. ISBN: 9780252725487.

Shosted, Ryan K. July 2006. "Correlating complexity: A typological approach". English (US). In: *Linguistic Typology* 10.1: 1–40. ISSN: 1430-0532. DOI: 10.1515/LINGTY.2006.001.

Trager, George L. 1955. *Language*. In: *Encyclopædia Britannica*. Vol. XIII: 695–702.

Wikimedia Foundation. 2021. *Wikipedia: Size of Wikipedia*. URL: https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia.

Winge, Johan. 2015. "Automatic annotation of Latin vowel length". Uppsala University. URL: https://cl.lingfil.uu.se/exarb/arch/winge2015.pdf.