# Quantifying the Diachronic Productivity of Irregular Verbal Patterns in Romance*

*Kevin Tang and Andrew Nevins*

**Abstract**

In this paper, we address the unproductivity of irregular verbal "L"-patterns in Portuguese, Italian and Spanish diachronically in a corpus linguistic study. Using openly available corpora, we answer two questions systematically: firstly whether the size of an active lexicon of a speaker/community remains constant, and secondly, whether the productivity of the regular verbal forms in the first conjugation *-ar(e)* increases over time and is a function of verb vocabulary size.

By running random sampling simulations on both large and small corpora from different sources for each language, we found a consistent increase, especially after 1750, in both verb vocabulary size and productivity of the regular verbal form *-ar(e)*. The results suggested that productivity of the regular verbal form is likely to be caused by the increase in verb vocabulary size, and as more new verbs come into a language, they will most likely fall into the first conjugation. This increase in the ratio of new verbs being assigned to the first conjugation caused the irregular forms in the second and third conjugations *-er(e)* and *-ir(e)* to become less productive over time. Finally, we speculate that the 1750 shift across all corpora is possibly caused by the industrial revolution which started around 1760.

*Keywords:* historical linguistics, productivity, irregular verbs, Romance languages

## 1 Introduction

In a number of Romance languages (we focus here on Portuguese, Italian and Spanish), the number of 'morphomic' verbs with the irregular 'L-pattern' (Maiden, 2005) between the 1st person singular form and the entire subjunctive seems no longer productive (Nevins & Rodrigues, 2012), although it was productive around 800 years ago.

(1)

| 'to say' | Indic | Sbj |
|---|---|---|
| 1sg | **dig**-o | **dig**-a |
| 2sg | diz-es | **dig**-as |
| 3sg | diz | **dig**-a |

Diachronically, the L-shape is essentially a consequence of the theme vowels that follow the stems causing palatalization. In the II/III conjugation, the 1sg.indic and sbj forms have in common a [+back] vowel, which enjoys the velar alternant, while the others have a [−back] vowel, with the palatal/coronal alternant.

Longer after the cessation of the process of palatalization in verb stems, this L-shaped pattern was apparently extended to verbs lacking a phonological reason for identity between the 1sg and sbj:

* We would like to thank Mark Liberman, Mark Aronoff, and Michael Becker for the initial suggestions on how to model verb vocabulary size diachronically, to Sean Wallis for providing critical comments on the methodology, to Joel Wallenberg for providing stimulating ideas and corpus suggestions and to Charles Yang for dialogue during the development of this paper.

(2)

| 'to hear' | Indic | Sbj |
|---|---|---|
| 1sg | **ouç**-o | **ouç**-a |
| 2sg | ouv-es | **ouç**-as |
| 3sg | ouv-e | **ouç**-a |

In an experimental study conducted across Portuguese, Spanish, and Italian, Nevins and Rodrigues (2012) found that this pattern is no longer productive for speakers furnished with partial paradigms of nonce words and asked to generalize to new inflectional forms; in fact, synchronically, speakers seem to prefer the opposite of the L-shaped pattern, seeking identity across persons or mood, instead of the morphosyntactically unnatural L-shape. The question, therefore, is why this L-shaped pattern was productive before but not now. Our hypothesis is that a number of irregular distributional patterns in the Romance verbal systems have disappeared from the language because the overall number of verbs in the language is larger now than it was 800 years ago.

To tackle this hypothesis, we examine two separate research questions in turn.

Question 1 – *Verb Vocabulary Size*: One might imagine that L-shaped verbs have ceased to become productive because they now represent a smaller proportion of the lexicon than they used to. Does verb vocabulary size would increase or stay constant diachronically; in another words, is the number of verbs in a speaker/community's active lexicon finite or stable or bounded over time?

Question 2 – *Productivity of ar-er-ir*: Suppose that the answer to Question 1 is that indeed the overall number of verbs in the language is larger now than it was 800 years ago. The verbal systems of these three Romance languages are organized into three conjugations, called the *ar-er-ir* conjugations (Italian uses *are-ere-ire*, but we adopt a consistent terminology here for conciseness). The L-shaped verbs are restricted to the latter two conjugations. As new verbs have come in to the language, are they imported to the *-ar* class, and as a result, do they gradually overshadow the *-er/-ir* verbs which have the L-shape?

Here we conducted a historical corpus study to answer these questions. First, we tested Question 1 on English as well as Portuguese, Italian and Spanish. The reason for testing English is to examine if the effect holds for only our Romance languages or languages in general, such as a Germanic language like English. For Question 2, we tested whether the productivity of *-ar*, the regular verb form, relative to *-er* and *-ir*, would increase or stay constant diachronically in Portuguese, Italian and Spanish. Finally, a correlation analysis was performed with the temporal trend of verb vocabulary size and that of productivity.

## 2 Data Sources

Only openly-accessible corpora were used in this study. This has the benefit of allowing a full-scaled modelling of the historical changes, as opposed to restricted queries through a web-interface which usually also imposes a search limit. Furthermore our work is open to validations and further development by other interested researchers.

### 2.1 English

Two historical corpora are available for English – *CLMET3.0* (Diller, De Smet, & Tyrkkö, 2011) and *Old Bailey* (Huber, Nissel, Maiwald, & Widlitzki, n.d.). We examined only CLMET3.0, a genre-balanced corpus (while Old Bailey is mainly restricted to spoken language in court trials), and the largest corpus of historical English (34 mil., cf. 22 mil. in Old Bailey).

*2.1.1 CLMET3.0.* The Corpus of Late Modern English Texts, version 3.0, contains 34 million words across the period 1710-1920 (divided into three 70-year sub-periods). The texts were written by native British English author, and the corpus restricts the number of texts per author to three or less, and is genre-balanced – narrative fiction, narrative non-fiction, drama, letters, treatises and miscellanea.

Automated Part-of-Speech labelling (POS-tagging) was done using EngTagger (Coburn, 2008). The accuracy of the tagging on this corpus was not evaluated. For early modern German, Scheible, Whitt, Durrell, and Bennett (2011) showed that an 'off-the-shelf' POS-tagger on their raw corpus has an accuracy of 69.6% and with regularised spelling, the accuracy was improved to 79.7%. Since English is a Germanic language, a reasonable estimate of accuracy would be around 70%.

## 2.2 Portuguese

Three open historical corpora are available for Portuguese – *Corpus do Português* (Davies & Ferreira, 2006), *Colonia* (Marcos & Martin, 2013) and *Tycho Brahe* (Galves & Pablo, 2010). In the present study, we only examined Corpus do Português and Colonia. Tycho Brahe was not examined as it is smallest of the three (half the size of Colonia) and has a shorter time-span.

Google Ngram corpus is not available for Portuguese. Another large available corpus is Corpus do Português, but since the full texts were not available, it was used only to estimate the productivity, and not the verb vocabulary size. Colonia is the only corpus of Portuguese which texts were fully available, however the size of the corpus is relatively small with only 5.1 million words, which is unlikely to be representative of the language; despite this drawback, it was used to model the verb vocabulary size as well as productivity.

*2.2.1 Corpus do Português.* Corpus do Português is a corpus containing 45 million words, spanning the 1300s to the 1900s, of which 10 million words are from the 1500s–1700s, and 15 million are from the 1800s–1900s. After 1700, the texts are evenly divided between Portugal and Brazil. The 1900s texts are evenly divided among spoken genres, fiction, newspapers, and academic. The corpus was POS-tagged and lemmatized, although the accuracy was not reported. (Davies & Ferreira, 2006). The corpus is only accessible via a web interface with POS-tagging information. It allows for regular expression searches, with the following **fixed** time-epochs: 1300s, 1400s, 1500s, 1600s, 1700s, 1800s and 1900s and more. The POS-tagger employed was a proprietary tagger that Michael Ferreira and Mark Davis developed.

*2.2.2 Colonia.* Colonia is a corpus containing 5.1 million words. The texts were written by Brazilian and European authors in a balanced proportion (52 Brazilian texts and 48 European texts) and divided into five sub-corpora by century. The time span is between the 16th to the early 20th century. The lemmas were semi-manually corrected.

The POS-tagging accuracy was not evaluated for these corpora and the spelling was not normalised corrected. For historical Italian (1200–1881), Pennacchiotti and Zanzotto (2008) showed that an average accuracy of 73.5% can be achieved, therefore an estimated accuracy would be around 73% for having not normalised the spelling for a Romance language (Scheible et al., 2011; Hendrickx & Marquilhas, 2011).

### 2.3 Italian

Two Italian corpora were examined: *Google Italian Ngram* (Lin et al., 2012) and *DiaCoris* (Onelli, Proietti, Seidenari, & Tamburini, 2006). DiaCoris was used for only for productivity estimation because the full text was not available.

*2.3.1 Google-Ngram:Italian.* Unigrams from the Google Ngram corpus of Italian were used, containing 40,288,810,817 words with the time span of 1550–2009. We included unigrams beginning with the letter "A" to "Z", and removed numbers, punctuations and miscellaneous items. We simulated the raw corpus by expanding the unigrams by count, and grouping them by year. The corpus has a POS-tagging accuracy of 95.6% (Lin et al., 2012).

*2.3.2 DiaCoris.* DiaCoris (Onelli et al., 2006) is a corpus of 20 million words, comprising written Italian texts produced between 1861 and 2001. It was designed to be a representative and well balanced sample of the Italian language, containing all the main events of recent Italian history such as the National Unification and the Second World War, and is sourced from the following genres: press, fiction, essayistic prose, legal-administrative prose and miscellanea. The time span of the corpus was split into four major periods, "After National Unification", "The Liberal Period", "Fascism", and "Post-fascism", each containing 5 million words, and thus resulting in a reasonably homogeneous corpus. At the moment, the corpus is only accessible via their web interface without POS-tagging information. It allows for regular expression searches, with the following **fixed** time-epochs, 1861–1900, 1901–1922, 1923–1945, 1946–1967 and 1968–2001, and the options of selecting individual sub-corpora.

### 2.4 Spanish

Two Spanish corpora were examined: *Google Spanish Ngram* (Lin et al., 2012) and *IMPACT-es* (Sánchez-Martínez, Martínez-Sempere, Ivars-Ribes, & Carrasco, 2013). IMPACT-es was used for only for productivity estimation, but not verb vocabulary estimation, because we have found that verb vocabulary estimations are more sensitive to the size of a corpus.

*2.4.1 Google Ngram:Spanish.* Unigrams from the Google Ngram corpus of Spanish were used, containing 83,967,471,303 words with the time-span of 1522 to 2009. We included unigrams beginning with the letter "A" to "Z", numbers, punctuations and miscellaneous items were removed. We simulated the raw corpus by expanding the unigrams by count, and grouping them by year. The corpus has a POS-tagging accuracy of 96.9% (Lin et al., 2012).

*2.4.2 IMPACT-es.* IMPACT-es is the only existing openly accessible historical corpus of Spanish (Sánchez-Martínez et al., 2013). It contains approximately 8 million words, from 107 Spanish texts first printed between 1481 and 1748. They cover a representative variety of creators and genres. It has two subset corpora, 6 million words come from the 21 Spanish documents in the ground-truth data set by IMPACT; the remaining 2 million words come from 86 texts provided by the Biblioteca Virtual Miguel de Cervantes digital library and are partially annotated (7%).

For our analyses, we used the latter smaller subcorpora because the larger subcorpus has not been normalised for spelling. Since the corpus is not POS-tagged, we POS-tagged the corpus using TreeTagger (Schmid, 1994). To increase the accuracy of the tagging, we utilised the annotated section of the corpus, which provided POS-tagging, lemmatisation and regularisation

of spelling for the 7% of the corpus which are mostly high frequency words. We used the regularized spelling whenever this was available, the information of the lemmatisation and POS-tagging were not used and were removed for re-tagging purposes.

The accuracy of using an 'off-the-shelf' POS-tagger on raw historical texts is unclear. For historical Spanish, Sánchez-Marco, Boleda, Fontana, and Domingo (2010) showed that an accuracy of 77.5% (POS-tagging) and 76.1% (Lemmatisation) could be achieved. A reasonable estimation of the accuracy would be around 75%.

## 3 Methods: Verb Vocabulary Size

We return to our research questions. The first is whether verb vocabulary size increases or stays constant diachronically; in another words, whether the number of verbs in a speaker/community's active lexicon is stable and/or bounded over time.

### 3.1 Simulations by Random Sampling

When comparing verb vocabulary size across different periods, we must consider the fact that the number of repertoire size is a function of sample size (Baayen, 2001), such that the larger the sample, the larger the estimated vocabulary size. For instance, in child language acquisition, when comparing parents' and children's verb vocabulary size (Ninio, 2011, Chapter 3), the parents' corpus is often much bigger than that of the children's. In order to avoid the aforementioned artefact, one technique is be to reduce the size of the bigger corpus to the smaller corpus by means of random sampling. Many random simulations must be obtained to estimate an average verb vocabulary size. We adopted this technique in this study, in which the corpora across all the periods/epochs are be reduced to the size of one of the smaller epochal corpora through random sampling. Using this method, 100 or 1,000 random simulations are conducted, to yield an average representation of changes in verb vocabulary size. The Google Ngram corpora were simulated only 100 times due to a time constraint imposed by the size of the corpora. The reason for not choosing the smallest epochal corpus period, is that the smallest epochal corpus can often be extremely small relative to the other epochal corpora, so to avoid losing a significant amount of data, and thus avoid undersampling, especially given that these diachronic corpora are already considerably smaller than synchronic corpora. Any epochs that cannot be matched to the fixed epochal corpus size were removed from the analyses.

### 3.2 Epoching

To estimate changes of vocabulary size over time, compared the changes every $N$ years. Three period sizes (epochs) were tested: 50, 25, and 10 years respectively. These sizes were selected based on plausible sizes of linguistic generations; smaller time windows would be unlikely to represent linguistic change, and larger time windows would potentially miss changes. In the current study, we used a fixed epoching window – that is, one with no overlap between epochs – e.g., Epoch(1700–1749), Epoch(1750–1799) etc. Any remainders from the epoching were removed from the analyses, e.g., if the whole time-span is 1700–1910 and the epoch size was 25, 1900–1910 would be the remainder from the epoching process.

For purposes of space, only the results for the 25 year epochs were shown (since we found that it was the most representative epochal size across all corpora, perhaps corresponding to the

time unit of a generation),with the exception of Corpus do Português and DiaCoris which have fixed epoch sizes limited by the online search interfaces.

### 3.3 Lemma estimation

To estimate the verb vocabulary size, the best approach would be to count the number of unique verb lemmas. However, most of corpora that we examined were POS-tagged, but not lemmatised, and even if they were lemmatised, many lemmas would not be found in a synchronic tagger, therefore an alternative way of estimation was needed. We used two verb forms: the infinitive form, and the (1st person singular) past tense form. These were used as separate estimates of the verb lemmas when the lemmas were not available. With English and the Romance languages, the infinitive form is arguably the most accurate representation of the lemma. The past tense form could also provide a highly representative estimation due to a likely bias of most texts (e.g., in reports and novels) containing more descriptions of the past than the present and future. More specifically with English, the past tense form does not vary with gender, person, and plurality, therefore this form was used only with English.

The Google Ngram corpora were syntactic parsed. The syntactic n-grams comprise of words (e.g., *burnt*), POS-annotated words (e.g., *burnt*_VERB), and POS tags (e.g., _VERB_). Only POS-annotated words were used in the analyses. They employed the universal part of speech tagset (Petrov, Das, & McDonald, 2011), containing only twelve POS tags: nouns, verbs, adjectives, adverbs, pronouns, determiners and articles, prepositions and postpositions, numerals, conjunctions, particles, punctuation marks, and other categories. The tagset does not make fine-grained distinctions between different verb forms, and therefore we limited the lemma estimation to the infinitive form, by using wildcard searches for words that end with *ar(e)*, *ir(e)*, *er(e)* with the verb tag.

## 4 Analyses: Verb Vocabulary Size

### 4.1 Simulation results: English, CLMET3.0

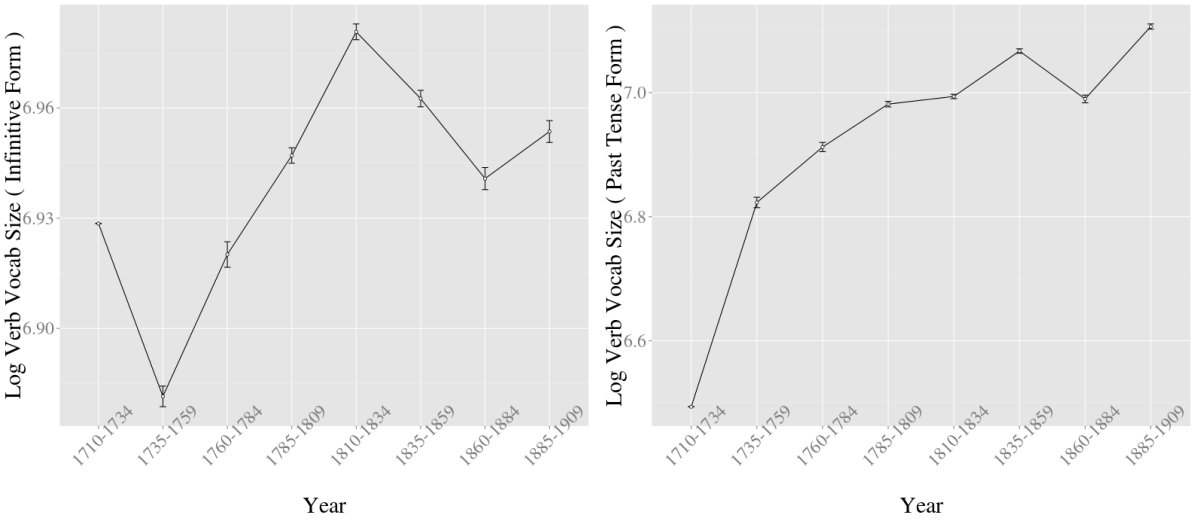
The diachronic corpus of English, CLMET3.0, showed a consistent increase of verb vocabulary size across a 200-year period (1710-1909), given 25-year epochs and lemma estimations for both infinitive (Figure 1a) and past tense (Figure 1b).

### 4.2 Simulation results: Portuguese, Colonia

Since Colonia has been lemmatised and manually corrected, no lemma estimation was needed for estimating verb vocabulary size. We conducted analyses both with the provided lemmas (Figure 2a) and based on the infinitive (Figure 2b). We found that the verb vocabulary size increases across a 400-year period of 1525–1924 based on 25-year epochs, with a sudden jump at the 1750–1774 epoch and continued increase thereafter.

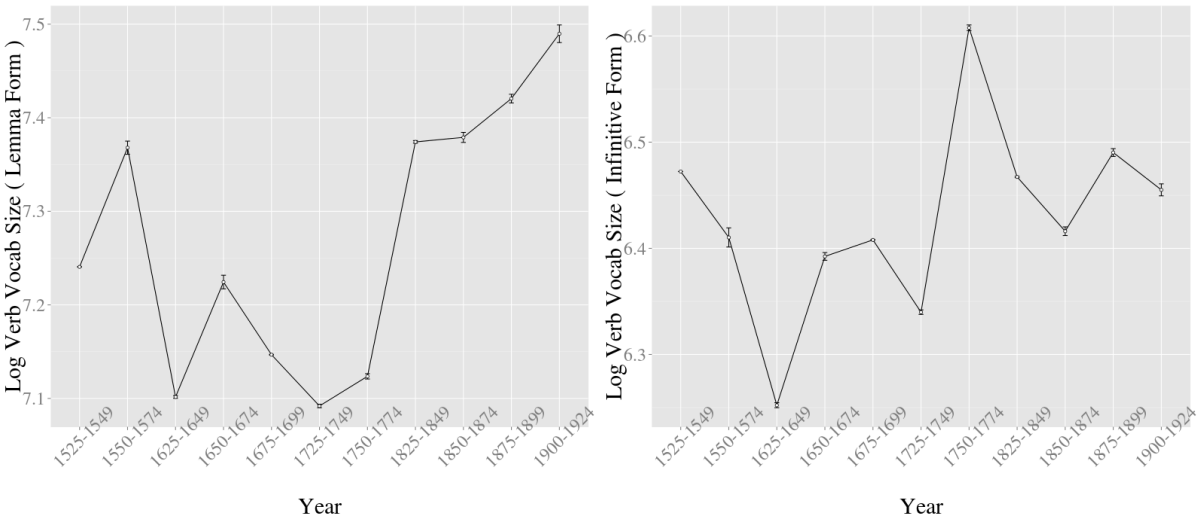
### 4.3 Simulation results: Italian, Google Ngram

The overall trend with the Google Italian Ngram corpus shows an increase in verb vocabulary size across a 450-year period (1550-1999) with 25-year epochs based on the infinitive as lemma (Figure 3), and a sudden jump at the 1750–1774 epoch, similarly to Portuguese. One of the epochs (1650-1674) appears to be an outlier.




(a) Lemma estimation: Infinitive tag (b) Lemma estimation: Past tense tag

Figure 1: 1,000 simulations of verb vocabulary size changes during 1710–1909; Language: English, Corpus: CLMET3.0, Lemma estimation: Infinitive and Past tense tags, Epoch size: 25 years, Epochal corpus size: 621,190




(a) Lemma estimation: None (b) Lemma estimation: *-ar/-er/-ir* with infinitive tag

Figure 2: 1,000 simulations of verb vocabulary size changes during 1525–1924; Language: Portuguese, Corpus: Colonia, Lemma estimation: None (using the lemmatised corpus) and *-ar/-er/-ir* with verb tag, Epoch size: 25 years, Epochal corpus size: 114,173

### 4.4 Simulation results: Spanish, Google Ngram

The overall trend with the Google Spanish Ngram corpus shows an increase in verb vocabulary size across the 475-year period from 1522–1996 in 25 year epochs, based on lemmatisation with the infinitive (Figure 4). There is a sudden jump at the 1722–1746 epoch, just as was found for Portuguese and Italian.

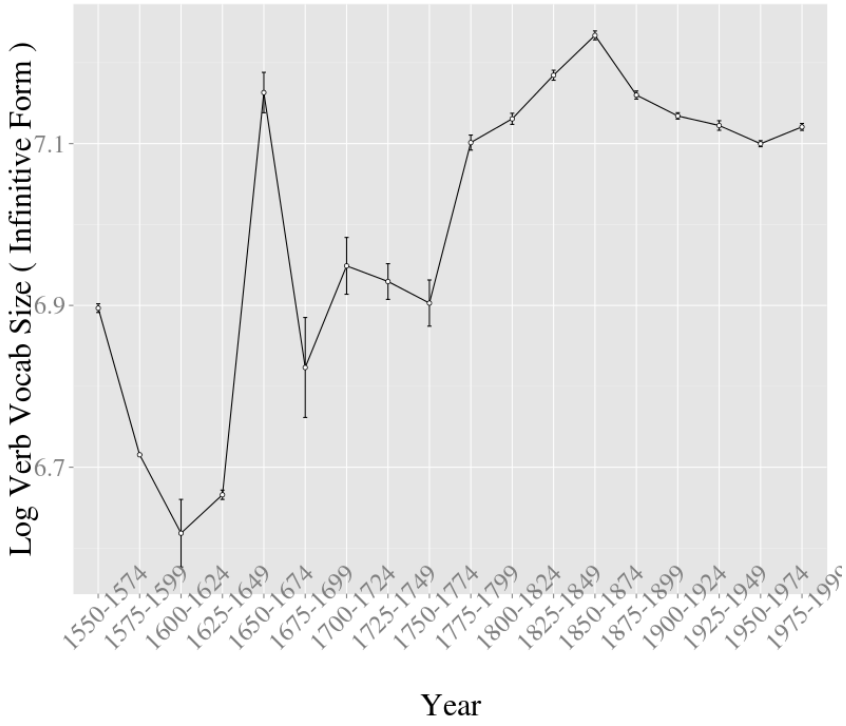



Figure 3: 100 simulations of verb vocabulary size changes during 1550–1999; Language: Italian, Corpus: Google Ngram, Lemma estimation: *-ar/-er/-ir* with verb tag, Epoch size: 25 years, Epochal corpus size: 633,911

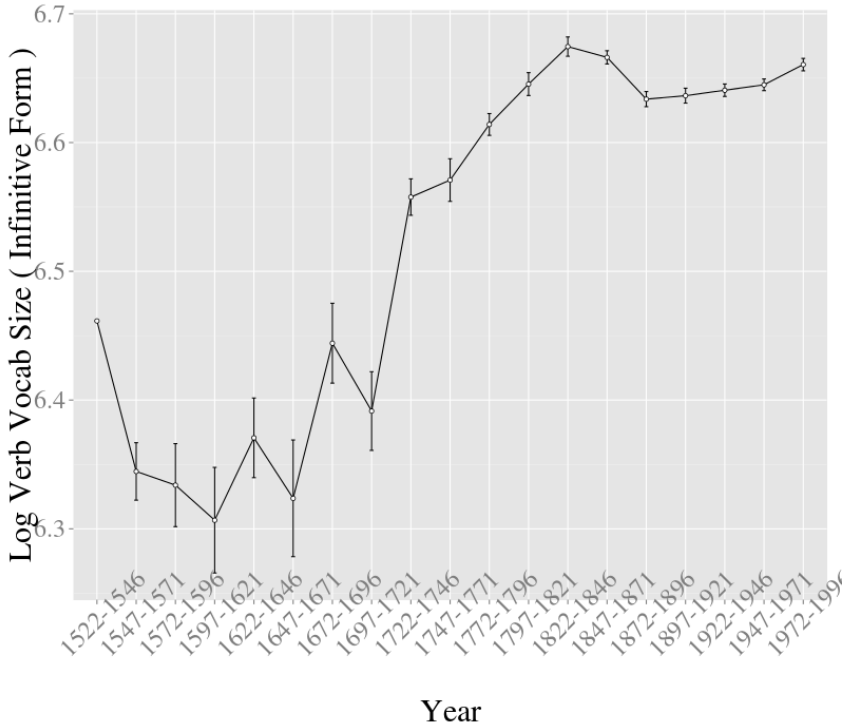



Figure 4: 100 simulations of verb vocabulary size changes during 1522–1996; Language: Spanish, Corpus: Google Ngram, Lemma estimation: *-ar/-er/-ir* with verb tag, Epoch size: 25 years, Epochal corpus size: 242,466

### 4.5 Interim Summary

We measured the increased in verb vocabulary size across the three Romance languages in question, as well as English, and found that all of them show an overall increase over time. Crucially this increase always measured with a fixed, equal vocabulary size overall for each epoch, based on the size of the epoch with one of the smallest available data, and submitted to a large set of random samples (100 or 1,000). The overall findings suggest that indeed, the number of verbs in these languages is on the increase over time, which relates to the hypothesis that the dwindling effect of productivity of the L-shaped verbs is due to their being overshadowed in the lexicon as more verbs come in (either through neologisms, loanwords, coinages, denominal derivation, or whatever means). Orthogonally to the question at hand, we found overall jumps in verb vocabulary size coinciding roughly with the time period of the Industrial Revolution in Europe.

## 5 Methods: Productivity of *ar-er-ir*

Having conducted the simulations across these corpora for the question of overall verb vocabulary size and found an increase, we turn to the more specific question of whether the three Romance languages show a change in productivity for their *-ar* and *-er/-ir* verb conjugations. This relates to the specific hypothesis that L-shaped verbs have lost their productivity not only because the overall number of verbs in the language is larger, but specifically because the *er-ir* conjugations, of which they are part, have decreased in productivity relative to the *-ar* class, which is where the majority of new verbs are placed.

### 5.1 Simulations by Random Sampling

Similar to the random sampling method in Section 3.1, in order to measure the productivity of each verb class, it is necessary to match the sizes of each epoch. The difference in the present case is that for each epoch, we matched the overall number of verbs instead of the overall number of words; previously we modelled the distribution of word types (e.g., verbs, nouns, etc.), but in this case, we modelled the distribution of verb types (*-ar, -er*, and *-ir*). This allows us to conduct a fair comparison of the distribution of the three verb types.

### 5.2 Productivity Estimation

*5.2.1* $\sum ar/(\sum er + \sum ir)$. By calculating the ratio of *-ar* versus *-er* plus *-ir*, we could estimate their relative productivity. If *-ar* were to become increasingly productive over time, then when a new verb enters the language, it should be more likely to fall under the *-ar* type, and the ratio would have an increasing trend over time.

*5.2.2 Yang's Productivity Estimate.* Yang (2005)'s tolerance principle was used to estimate the productivity of *-ar* verbs. The theorem states $M \approx N/ln(N)$, where $M$ is the number of exceptions/irregular forms and $N$ is the number of verbs.

We estimated the irregular form for $M$ using *-er* and *-ir* forms. We understand that this is an overestimation, as not *-er/-ir* verbs are irregular; however, the majority of irregular verbs are in the *-er/-ir* class, and this thus provides a way to examine the distribution of irregular verbs without a diachronic, language-specific list. The productivity values that we report should not

be interpreted directly, only the *relative* productivity across time is relevant for considering the research question at hand.

Using Yang's formula, for every given *M* (per period), we calculate the minimal number of verbs required for the regular rule of *-ar* to be safe, by solving *N*. Since our *M* is always an overestimate, our *N* (minimal number of verbs) will also be an overestimate. The productivity of *-ar* is therefore

$$1 - \frac{(\text{Minimal Number of Verbs} - \text{Total Number of Verbs})}{\text{Minimal Number of Verbs}}$$

Our analyses showed that the two methods of estimation yield a nearly identical trend, therefore only the results with the former method by $\sum ar/(\sum er + \sum ir)$ were shown below.

## 6 Analyses: Productivity of *ar-er-ir*

### 6.1 Simulation results: Portuguese, Corpus do Português

Although Corpus do Português was lemmatised, only the tagging information was used. We extracted all the verb-tagged words with the following three wildcards, "*ar", "*er", and "*ir". The overall trend, shown in Figure 5c shows a stable increase in productivity of *-ar* from 1300 to 1900, with fixed epochs of 100 years (as provided/limited by the query interface).

### 6.2 Simulation results: Portuguese, Colonia

The overall trend is less clear than the Corpus do Português, based on 25-year epochs with the lemmatised version (Figure 5a) and based on the infinitive (Figure 5b). However, there is a steady increase in productivity of *-ar* after the 1750–1774 epoch, just as with the trend of verb vocabulary size.

### 6.3 Simulation results: Italian, Google Ngram

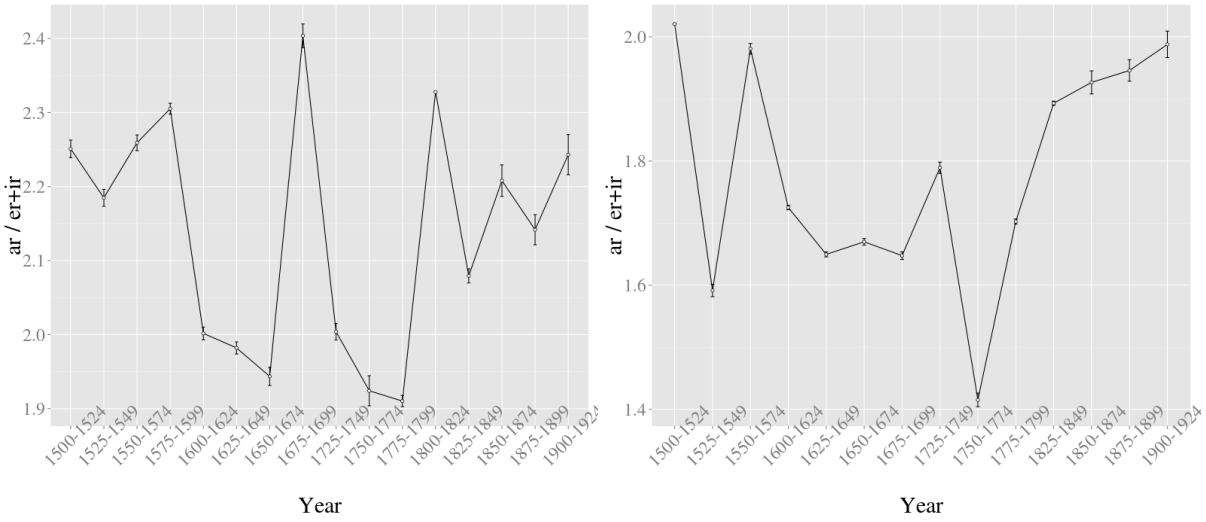
For Italian, the verb types are *-are, -ere* and *-ire*. across a 425-year period (1550–1974) with 25-year epochs based on the infinitive (Figure 6a), and a sudden jump at the 1750-1774 epoch, just as with the trend of verb vocabulary size. One of the epochs appeared to be an outlier at 1650–1674.

### 6.4 Simulation results: Italian, DiaCoris

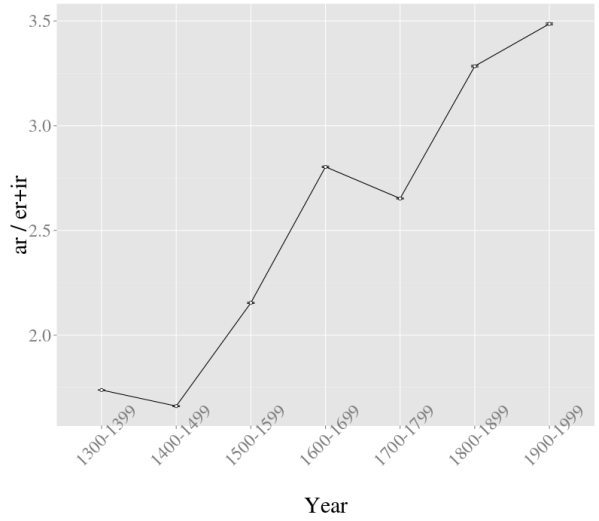
DiaCoris is not tagged; we therefore extracted all the verbs with wild-cards "*are", "*ere", and "*ire", across the fixed epochs. It was not possible to match the epoch sizes with the web interface, which are in the range of 22–40 years. The overall trend again shows an increase in productivity of *-ar* from 1861 to 2001 (Figure 6b), which matches the trend with Google Italian Ngram in the period. This suggests that the trend we found is unlikely to be an artefact of corpus selection.

### 6.5 Simulation results: Spanish, Google Ngram

The overall trend with the Google Spanish Ngram corpus shows an increase in productivity of *-ar* (Figure 7a), but a sudden jump at the 1747–1771 epoch, just as the trend of verb vocabulary size.




(a) Lemma estimation: None (b) Lemma estimation: *-ar/-er/-ir* with infinitive tag

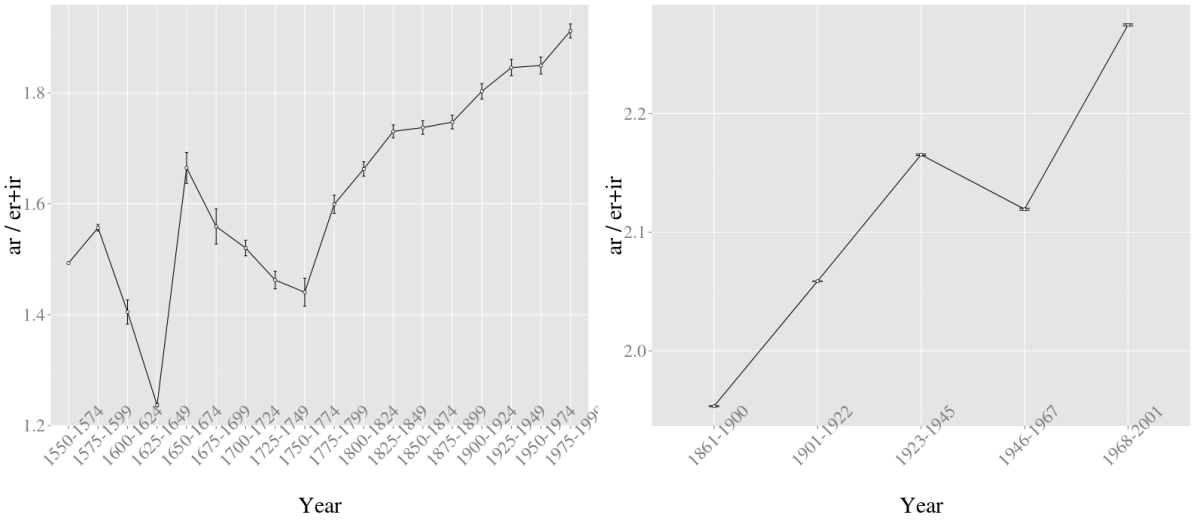



(c) Lemma estimation: *-ar/-er/-ir* with verb tag

Figure 5: 1,000 simulations of productivity changes of Portuguese (a & b) Colonia, 1525–1924, Epoch size: 25 years, Epochal corpus size of target verb tokens: 1,252 and 1,285 respectively and (c) DiaCoris, 1300–1999, Epoch size: 100 years, Epochal corpus size of target verb tokens: 41,751

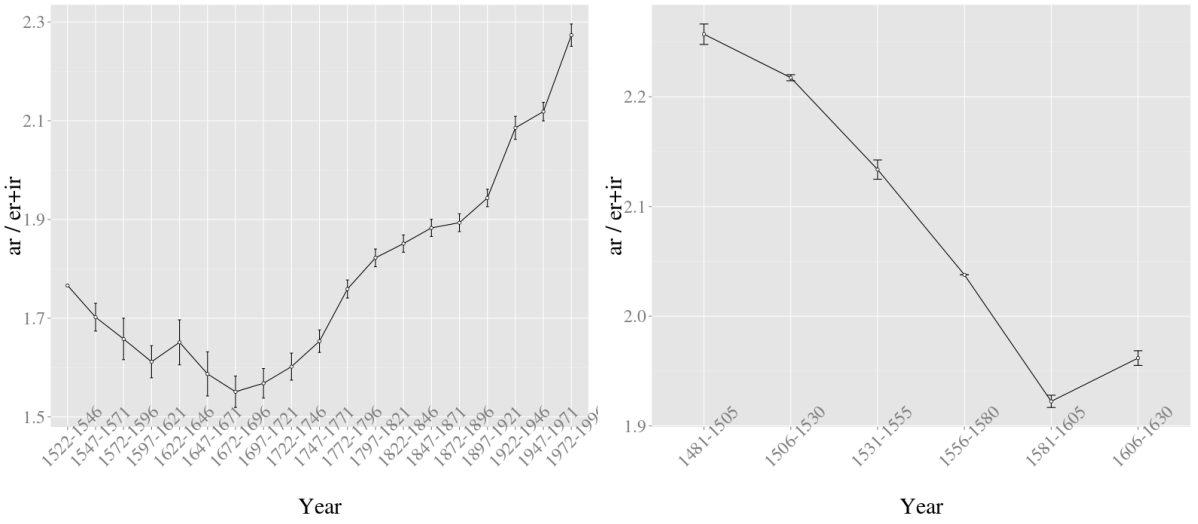
### 6.6 Simulation results: Spanish, IMPACT-es

The IMPACT-es corpus does not extend over the same historical range as the Google Ngram Spanish corpus, but we wanted to conduct a validation of the trend found in the 1481–1630 time period. The overall trend shows an decrease in productivity of *-ar* (Figure 7b) across 25-year epochs, which matches the trend with Google Spanish Ngram in the period. While the period of interest is arguably well after 1630, these findings nonetheless confirm the calculations possible during comparable periods using different corpora.




(a) Lemma estimation: *-ar/-er/-ir* with verb tag (b) Lemma estimation: *-ar/-er/-ir* without verb tag

Figure 6: 100/1,000 simulations of productivity changes of Italian (a) Google Italian Ngram, 1550–1999, Epoch size: 25 years, Epochal corpus size of target verb tokens: 5,456 and (b) DiaCoris, 1861–2001, Epoch size: Predefined years (average 28.5 years), Epochal corpus size of target verb tokens: 109,000




(a) Lemma estimation: *-ar/-er/-ir* with verb tag (b) Lemma estimation: *-ar/-er/-ir* with infinitive tag

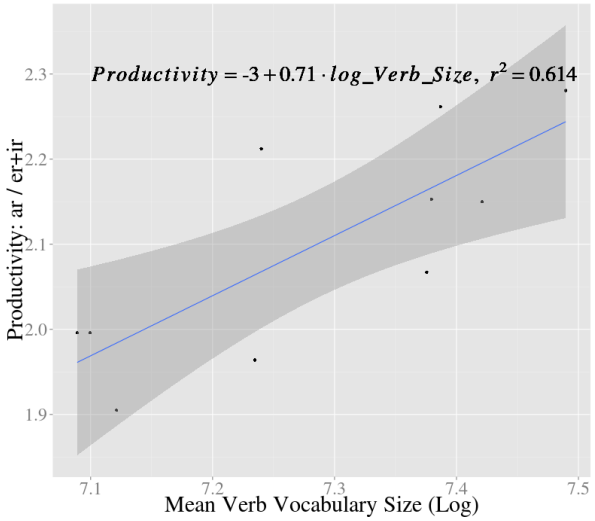
Figure 7: 100/1,000 simulations of productivity changes of Spanish (a) Google Spanish Ngram, 1522–1996, Epoch size: 25 years, Epochal corpus size of target verb tokens: 2,646 and (b) IMPACT-es, 1481–1630, Epoch size: 25 years, Epochal corpus size of target verb tokens: 1,554

## 7 Relationship between Verb vocabulary size and Productivity

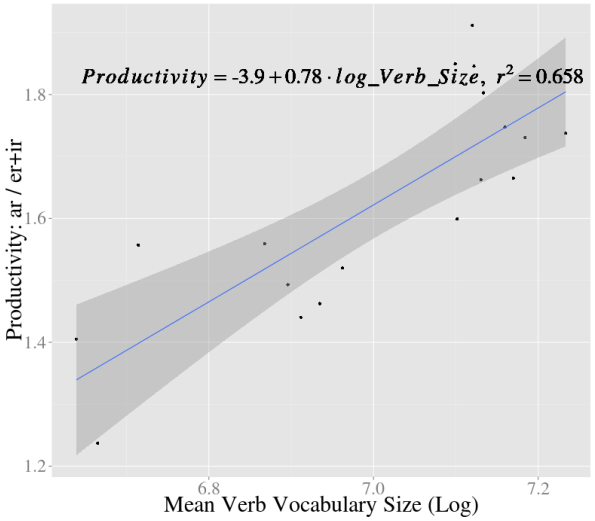
In order to establish whether the trend of productivity is related to that of verb vocabulary size or not, a correlation analysis was performed on Colonia, Google Italian Ngram and Google Spanish Ngram. The mean value (of all the random samples) was used.

There was a strong and significant correlation between verb vocabulary size and productivity – 1) Portuguese (Colonia) (Figure 8a): $r(8) = 0.78$ , $p = 0.0071$; 2) Italian (Google

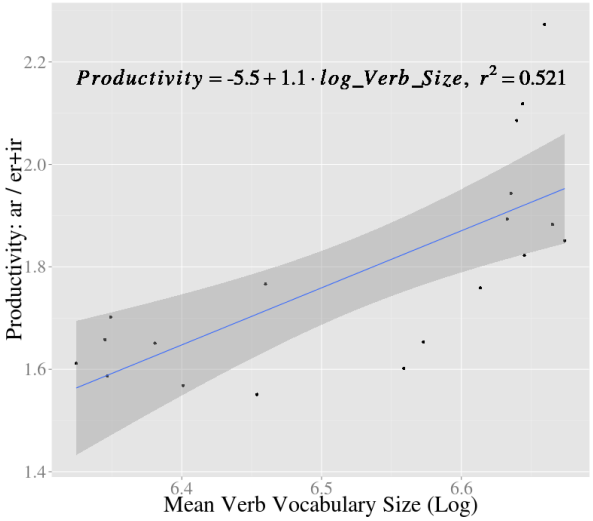
Ngram) (Figure 8b): $r(16) = 0.81$ , $p = 4.461e\text{-}05$; 3) Spanish (Google Ngram) (Figure 8c): $r(17) = 0.72$ , $p = 0.00045$. Given the small corpus size of Colonia, it was not expected to be very revealing, nevertheless the correlation was significant after removing an outlier epoch 1675–1699 (see Figure 5b). Furthermore by overlaying the trends, the correlation between verb vocabulary size and productivity is highly transparent. (Figure 9, 10 and 11).




(a) Portuguese




(b) Italian




(c) Spanish

Figure 8: Relationship between verb vocabulary size and productivity; (a) Portuguese (Corpus: Colonia), (b) Italian (Corpus: Google Ngram), (c) Spanish (Corpus: Google Ngram)

## 8 Statistical evaluation of the changepoint of verb vocabulary growth

Thus far, we have visually observed that there is a sudden increase in both verb vocabulary size and productivity of *-ar* at around or slightly after 1750. A changepoint analysis was conducted to statistically quantify this observation. The R package *changepoint* (Killick & Eckley, 2011), was used. Changepoint detection estimates the point(s) at which the statistical properties of a sequence of observations change. On the whole, there are two kinds of algorithms: single

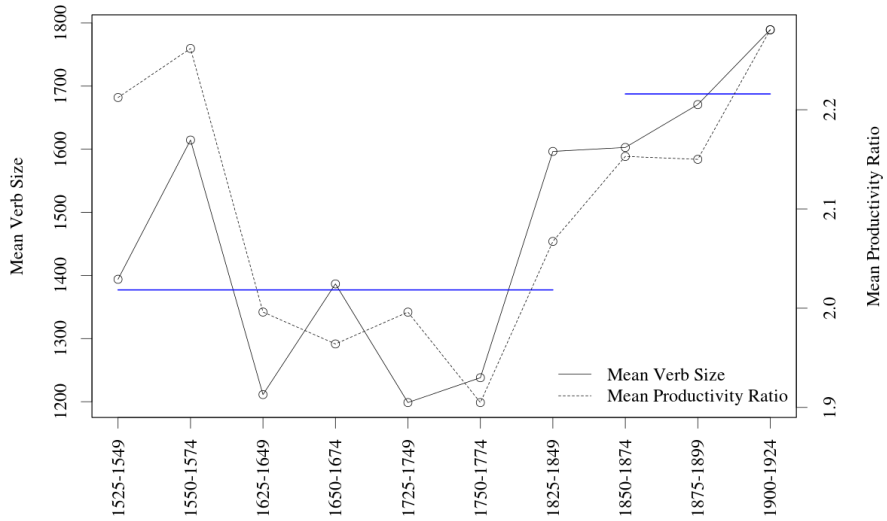



Figure 9: Temporal trends of verb vocabulary size and productivity with changepoint analysis; Language: Portuguese, Corpus: Colonia

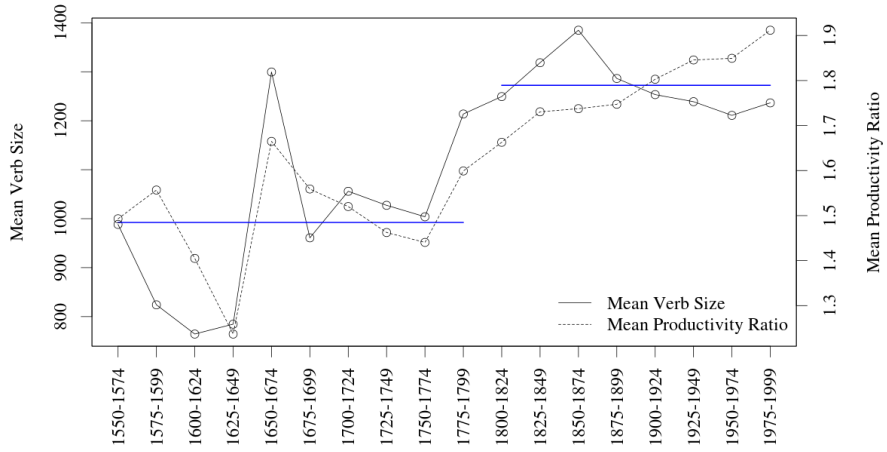



Figure 10: Temporal trends of verb vocabulary size and productivity with changepoint analysis; Language: Italian, Corpus: Google Italian Ngram

or multiple changepoint detection. Due to the relatively small number of epochs, the multiple changepoint detection method is not meaningful, since every epoch would be treated a changepoint, and we therefore employed the single changepoint detection method, which allows at most one change in the detection. Furthermore, since our data violate the normal distribution assumption, we selected the Cumulative Sum (CUSUM) test statistics (Page, 1954) which have no distributional assumptions.

We applied the single change detection method with CUSUM statistics to the mean values of the verb vocabulary size simulations for the Romance languages. English was excluded from this analyses due to limited epochal range. We found there was a statistical significant change in each of the corpora, and the epochs at which this took place are as followed: In the corresponding plots, a change in mean is indicated by horizontal lines depicting the mean value

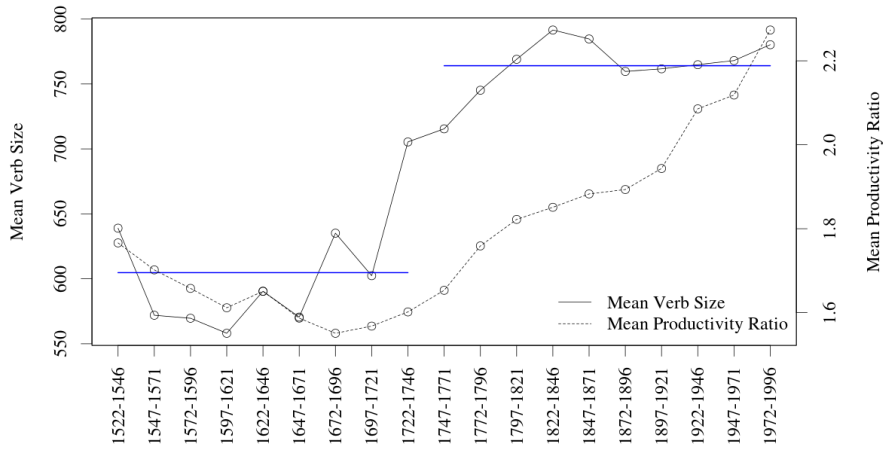



Figure 11: Temporal trends of verb vocabulary size and productivity with changepoint analysis; Language: Spanish, Corpus: Google Spanish Ngram

in different segments, where the disjunctures are the changepoints.

- Portuguese (Colonia) (Figure 9): 1825–1849
- Italian (Google Ngram) (Figure 10): 1775–1799
- Spanish (Google Ngram) (Figure 11): 1722–1746

Although the epoch where the change occurred are not identical, they are all clustered within less than one hundred years around 1750. For the Colonia corpus, the result is the same with or without taking out the outlier epoch 1675–1699. While the reason why verb size shows a sudden growth in this period across all three of these languages remains to be found, we speculate that the change is related to the Industrial Revolution, which greatly changed society not only in terms of technology but also in terms of increased travel, mobility, education, and health/lifespan, and one never knows whether it was these secondary/indirect factors that had/have the most influence on linguistic change and vocabulary growth.

Further studies are needed to examine this Industrial Revolution hypothesis, perhaps by comparing the languages spoken in countries with different degrees of effects from the Revolution. Studies have suggested that the Revolution began in Great Britain and did not take full effect in the Netherlands until the last third of the 19th century (Mokyr, 2000; Allen, 2009); if so, then we would expect to see a sudden change around 1750 for English, and a late or perhaps no sudden change for Dutch. In principle, this work could be related to the comparison of verb vocabulary size with a number of economic and technological changes in different language communities.

## 9 Artefact considerations

In this section, we will consider a range of potential artefacts that might have affected, or indeed provide alternate explanations, for our pattern of results. A number of these reflect considerations directly related to limitations in the resources available for these Romance languages at present.

### 9.1 Corpus representativeness

One major criticism in corpus linguistics is the representativeness of the corpora. Many have addressed how to achieve good representativeness in corpus design (Atkins, Clear, & Ostler, 1992; Biber, 1993). Representativeness often refers to how much a sample contains the same range of variability in a population. Two main kinds of variability are situational (the type of text) and linguistic. If a corpus fails to represent the range of texts in the target population, it will therefore fail to represent the range of linguistic distributions. Just like experimental controls, corpus compilers aim to control for many variables, such as the genres, the number of words per document, the number of documents per document type, the gender of the authors, the register of the documents, the number of authors and many more.

In diachronic corpora, many variables cannot be controlled for, mainly due to the lack of data. In most cases, a trade-off between corpus size and representativeness is made by the compilers, whose aim is often simply to include as many documents as possible, without controlling for the aforementioned variables.

These corpora are samples of the language of a small group of literate writers. The number of writers and documents, as well as the genre-range that is sampled will undoubtedly increase over time. Therefore, one possible corpus artefact which could explain our verb vocabulary size results could be that the growth in number of verbs is a result of a widening social milieu of literacy and genres.

Random re-sampling will not completely address this, since the problem lies with the initial distribution of the documents. The solution will have to be re-sampling of the source of the corpora by matching the sub-distributions of genres, the number of authors and more.

In this paper, we used a wide range of corpora, and in most cases, we studied all the openly-available corpora. Most of the corpora were pre-compiled and processed such as the Google Ngram corpora, or available only via web search-engines such as Corpus do Português and DiaCoris, and the unavailability of the full texts made it impossible at present to take into account of various corpus controls by carefully resampling the corpora. In the case of Colonia and CLMET3.0, the distributions are known and have been partially matched by genres and the number of authors. Furthermore, Google Ngram for English – an "unbalanced" corpus which included as many books as possible – will be used in the future to examine the verb vocabulary estimation, and if the result is again consistent with that of CLMET3.0, a balanced corpus, this would strengthen our overall findings and weaken representativeness as an artefact.

However, given that we analysed multiple corpora, with two corpora per language group (with the exception of English), we deemed that corpus representativeness as an artefact is unlikely to explain the consistency of our findings.

### 9.2 Tagging accuracy and consistency

Two possible artefacts lie with the POS taggers, namely the accuracy and consistency across taggers, which in turn affect our estimations of lemmas with a stronger effect in verb vocabulary estimation, and a relatively minor effect in the productivity estimation.

One could argue that the taggers' accuracy improves over time and therefore more verbs were correctly identified and thus more verbs were found. For historical Italian, Pennacchiotti and Zanzotto (2008) tested the accuracy of a synchronic tagger for Italian on diachronic data, and they did not find a consistent increase in accuracy across time; while for historical German, Scheible et al. (2011) showed that there is an increase in accuracy across time if the texts are unprocessed, but that there is no such increase when the texts are standardised for spelling. These

seem to suggest that the consistent increase might exist with languages with radical spelling changes such as German, but less so for Romance languages. Nevertheless, this gives us reason to think that spelling normalisation would be one way to avoid this artefact,

Table 1 summarizes the information about spelling normalisation, tagging and lemmatisation and the accuracy of all the corpora. It is clear that spelling normalisation was absent in all but the CLMET3.0 corpus, so there is good reason to suspect that the accuracy would indeed increase over time. However, the Google Ngram corpora have employed a word clustering technique which circumvented the issue with spelling variants affecting POS-tagging accuracy. The technique is to cluster words based on their distributional properties, and use them as features in their POS tagger. This allows unknown words (words that are spelled differently, incorrectly OCR-ed or simply rare) to be tagged correctly because they share similar co-occurrence patterns with the known/correctly spelling versions (Lin et al., 2012). In sum, the results on the verb vocabulary size and productivity using the Google Ngram corpora (Italian and Spanish) should not suffer from this artefact.

The second potential artefact due to taggers is their relative consistency. Each tagger might have high accuracy but ideally the same model of tagger (trained on different language data) should be applied for all the corpora in question. Clearly this is not possible due to a) the lack of historical taggers (with the exception of Sánchez-Marco et al. (2010)'s historical FreeLing tagger for Spanish, though not applied in this study) and b) the availability of the source texts.

The tagger-related artefacts should not be a major issue for our productivity estimation. In the productivity estimation, we concerned only with the distribution of verb classes, independent from the overall number of verbs (given that we held the epochal size in our simulation constant across time). However, the artefacts could be an issue if the tagging accuracy is uneven amongst the verb classes; that is, the diachronic accuracy of tagging *-ar* is different from that of tagging *-er/-ir*. This could bias our results if:

- The accuracy of tagging *-ar* is stable across time, while that of tagging *-er/-ir* decreases;
- The accuracy of tagging *-ar* increases, while that of tagging *-er/-ir* is stable;
- The accuracy of tagging *-ar* increases, while that of tagging *-er/-ir* decreases.

However, there is no evidence that we know of that might indicate the tagging accuracy is uneven amongst the verb classes.

Furthermore, the lemmatised corpora could reduce the effect of tagging accuracy on productivity estimation. Although the POS tagging accuracy for Corpus do Português and Colonia were not verified, the lemmas were. With Corpus do Português, the lemmatisation was done automatically as well as manually whenever needed (i.e. when a lemma cannot be identified). More specifically, in the earlier years the corpus was heavily annotated manually which is particularly reassuring as those are the periods where lemmatisation would fail most. Similarly with Colonia, the lemmas were semi-manually verified. Nonetheless, the incorrectly tagged words would still have incorrect lemmas, and as a result, while the lemmas are not totally reliable, it is a way to reduce the effect of this artefact. Therefore, the productivity results using the lemmatised Colonia and Corpus do Português *without* POS-tags should suffer less from this artefact. (Recall that the lemmatised Corpus do Português was not used and will be included in the future development of this study) Finally, the Italian DiaCoris corpus was not tagged and was searched using wild-cards, and yet the result was consistent with those from other corpora.

A few potential solutions besides retagging are possible, a) searching with wildcards, and b) checking for false positives. Firstly, by searching with wildcards disregarding the tags, we could see if we could arrive at the same conclusion, allowing us to triangulate our results.

| Language | Corpus | Spelling Normalised | Tagged | Accuracy | Lemmatised | Verified |
|---|---|---|---|---|---|---|
| English | CLMET3.0 | Yes | Yes | Unknown Est. 70% | No | N/A |
| Portuguese | Corpus do Português | No | Yes | Unknown Est. 73% | Yes | Yes |
| Portuguese | Colonia | No | Yes | Unknown Est. 73% | Yes | Yes |
| Italian | Google Ngram | No | Yes | 95.6% | No | N/A |
| Italian | DiaCoris | No | No | N/A | No | N/A |
| Spanish | Google Ngram | No | Yes | 96.9% | No | N/A |
| Spanish | IMPACT-es | Partially (7%) | Yes | Unknown Est. 75% | Partially (7%) | Yes |

Table 1: Corpus summary

However one could argue that it is possible (but unlikely) that there was an increase of non-verb lexical items with the ending with *-ar*/*-er*/*-ir*. Secondly, instead of wildcard searching, we could extract all the words with a tag that is not a verb, and end with *-ar*, *-er* or *-ir*; we then could manually check for how many of these allegedly non-verbs are verbs, and whether these false positives also increases over time. The potential solution with the wildcards is more feasible and preferable than that with the false positives. This is because a) the latter will require researchers with specializations in historical Spanish, Italian and Portuguese, and b) arguably the manual tagging accuracy could also be an artefact, as the more recent forms of the languages are better documented than the more historical forms, and therefore more accurately tagged. The wildcard solution will be employed in the future development of this study.

In the preceding text, we have discussed the possible tagging artefacts on verb vocabulary size and productivity, and we proposed and conducted some of the solutions. For English, CLMET3.0 is normalised for spelling which should remove the tagging bias just as the study of historical German (Scheible et al., 2011), and yet we still saw an increase in verb vocabulary size. For Italian and Spanish, the Google Ngram corpora should not suffer from these tagging biases in both verb vocabulary size and productivity due to their unique clustering technique. For Italian, the results using wildcards (therefore not affected by tagging biases) on DiaCoris showed a consistent increase just as the results from other corpora. For Portuguese, using the lemmatised Colonia, the results are again consistent. Jointly considering both the steps taken to address these artefacts and the consistent outcome, the artefacts from tagging are unlikely to be able to explain all of our findings. Further work such as wildcard searching will be conducted to strengthen this conclusion.

## 10 Conclusion

In this paper, we investigated the possible cause for the unproductivity of irregular verbs in Portuguese, Italian and Spanish.

Firstly, we analysed the change in verb vocabulary size across time of English, Portuguese, Italian and Spanish. All languages show a consistent increase in verb vocabulary size, suggesting the number of verbs (or perhaps words in general) in a speaker/community's active lexicon is not finite or bounded over time. Secondly, we analysed the productivity of *-ar*, the regular verb form, relative to *-er* and *-ir* using two productivity estimations, namely *-ar/(-ir+-er)* and Yang's productivity estimate. We found that again there is an increase in productivity of *-ar* dia-

chronically across all three languages. Thirdly, our correlation analyses showed that the three trends are strongly correlated with $r$ in the range of 0.7–0.8 and $p < 0.007$.

These findings together suggest that when a new verb enters the language, it is mostly allocated to the verb type *-ar*, and over time this overcomes the salience of the irregular verb forms *-er* and *-ir*, rendering the L-shaped pattern synchronically unproductive.

Finally, we observed a sudden increase in verb vocabulary size (therefore productivity) at around 1750 across the three Romance languages, and this was confirmed by an objective changepoint statistical analysis. The analyses showed that the range at which the sudden jump happened is 1722–1849. This led us to speculate that the reason for this sudden jump in the lexicons for these languages is tempting to relate to the Industrial Revolution.

## References


Allen, R. C. (2009). *The British industrial revolution in global perspective*. Cambridge University Press.

Atkins, S., Clear, J., & Ostler, N. (1992). Corpus design criteria. *Literary and Linguistic Computing*, *7*(1), 1–16.

Baayen, R. (2001). *Word frequency distributions*. MIT Press.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, *8*(4), 243–257.

Coburn, A. (2008). Lingua::EN::Tagger - search.cpan.org. Retrieved from http://search.cpan.org/perldoc?Lingua::EN::Tagger

Davies, M., & Ferreira, M. (2006). Corpus do Português (45 million words, 1300s-1900s). *National Endowment for the humanities*.

Diller, H., De Smet, H., & Tyrkkö, J. (2011). A European database of descriptors of English electronic texts. *The European English Messenger*, *19*, 21–35.

Galves, C., & Pablo, F. (2010). Tycho Brahe Parsed Corpus of Historical Portuguese. Retrieved from http://www.tycho.iel.unicamp.br/~tycho/corpus/en/index.html

Hendrickx, I., & Marquilhas, R. (2011). From old texts to modern spellings: an experiment in automatic normalisation. *JLCL*, *26*(2), 65–76.

Huber, M., Nissel, M., Maiwald, P., & Widlitzki, B. (n.d.). The Old Bailey corpus. spoken English in the 18th and 19th centuries. Retrieved September 24, 2013, from www.uni-giessen.de/oldbaileycorpus

Killick, R., & Eckley, I. A. (2011). Changepoint: an R package for changepoint analysis. *Lancaster University*.

Lin, Y., Michel, J., Aiden, E., Orwant, J., Brockman, W., & Petrov, S. (2012). Syntactic annotations for the Google books ngram corpus. In *Proceedings of the ACL 2012 system demonstrations* (pp. 169–174). Association for Computational Linguistics.

Maiden, M. (2005). Morphological autonomy and diachrony. In *Yearbook of morphology 2004* (pp. 137–175). Springer.

Marcos, Z., & Martin, B. (2013). Colonia: corpus of historical Portuguese. In Z. Marcos, & D. Sascha (Eds.), *Special volume on non-standard data sources in corpus-based research* (Vol. 5). ZSM Studien. Shaker.

Mokyr, J. (2000). The industrial revolution and the Netherlands: why did it not happen? *De Economist*, *148*(4), 503–520.

Nevins, A., & Rodrigues, C. (2012). Naturalness biases, 'Morphomes', and the Romance First Person Singular. Retrieved from http://ling.auf.net/lingbuzz/001469

Ninio, A. (2011). *Syntactic development, its input and output*. Oxford University Press.

Onelli, C., Proietti, D., Seidenari, C., & Tamburini, F. (2006). The DiaCORIS project: a diachronic corpus of written Italian. In *Proceedings of LREC-2006, the fifth international conference on language resources and evaluation* (pp. 1212–1215).

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, *41*(1/2), 100–115.

Pennacchiotti, M., & Zanzotto, F. (2008). Natural language processing across time: an empirical investigation on Italian. In *Advances in natural language processing* (pp. 371–382). Springer.

Petrov, S., Das, D., & McDonald, R. (2011). A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*.

Sánchez-Marco, C., Boleda, G., Fontana, J., & Domingo, J. (2010, May). Annotation and representation of a diachronic corpus of Spanish. In N. C. ( Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, …, & D. Tapias (Eds.), *Proceedings of the seventh international conference on language resources and evaluation (lrec 10)*. Valletta, Malta: European Language Resources Association (ELRA).

Sánchez-Martínez, F., Martínez-Sempere, I., Ivars-Ribes, X., & Carrasco, R. (2013). An open diachronic corpus of historical Spanish: annotation criteria and automatic modernisation of spelling. *arXiv preprint arXiv:1306.3692*.

Scheible, S., Whitt, R., Durrell, M., & Bennett, P. (2011). Evaluating an 'off-the-shelf' pos-tagger on early modern German text. In *Proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences, and humanities* (pp. 19–23). Association for Computational Linguistics.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of international conference on new methods in language processing* (Vol. 12, pp. 44–49). Manchester, UK.

Yang, C. (2005). On productivity. *Linguistic Variation Yearbook*, *5*(1), 265–302.