
This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.

Rodriguez-Puente, Paula; Säily, Tanja; Suomela, Jukka

New methods for analysing diachronic suffix competition across registers: How -ity gained ground on -ness in Early Modern English

Published in:
International Journal of Corpus Linguistics

DOI:
[10.1075/ijcl.22014.rod](https://doi.org/10.1075/ijcl.22014.rod)

Published: 20/10/2022

Document Version
Publisher's PDF, also known as Version of record

Published under the following license:
CC BY

Please cite the original version:
Rodriguez-Puente, P., Säily, T., & Suomela, J. (2022). New methods for analysing diachronic suffix competition across registers: How -ity gained ground on -ness in Early Modern English. *International Journal of Corpus Linguistics*, 27(4), 506-528. <https://doi.org/10.1075/ijcl.22014.rod>

This material is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the repository collections is not permitted, except that material may be duplicated by you for your research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone who is not an authorised user.

New methods for analysing diachronic suffix competition across registers

How *-ity* gained ground on *-ness* in Early Modern English

Paula Rodríguez-Puente, Tanja Säily, and Jukka Suomela
University of Oviedo | University of Helsinki | Aalto University

This paper tracks stylistic variation in the use of two roughly synonymous suffixes, the Romance *-ity* and the native *-ness*, during the Early Modern English period. We seek to verify from a statistical viewpoint the claims of Rodríguez-Puente (2020), who reports on a decrease of *-ness* in favour of *-ity* in registers representative of the speech-written and formal-informal continua at that time. To this end, we develop new methods of statistical and visual analysis that enable diachronic comparisons of competing processes across subcorpora, building upon an earlier method by Säily and Suomela (2009). Our results confirm that *-ity* gained ground first in written registers and then spread towards speech-related registers, and we are able to time this change more accurately thanks to a novel periodisation. We also provide strong statistical support indicating that the proportion of *-ity* was significantly higher in legal registers than in other registers.

Keywords: derivational morphology, Early Modern English, productivity, cross-register analysis, statistical analysis

1. Introduction

The Early Modern English period (EModE) was a crucial time in the expansion of the English vocabulary, not only as a result of large-scale borrowing but also of the highly productive use of word formation processes, which was greatly affected by the foreign influences of the time and the growing demands of the developing standard language (Nevalainen, 1999: 332–333, 336–337). Intensive borrowing in Middle English (ME) led to a decline of the native affixal system inherited from Old English (OE; Romaine, 1985: 461–462), but word-formation patterns were still irregular at the beginning of the EModE period, and freedom of choice



in affix usage resulted in a large number of doublets or parallel derivatives, such as *frequency* and *frequentness* (Nevalainen, 1999: 334). It is, however, hard to determine whether these doublets are exact synonyms in usage and meaning, even from the synchronic point of view (see Riddle, 1985; Dalton-Puffer, 1996: 126–130).¹

This paper examines stylistic variation in the use of two roughly synonymous suffixes, the Romance *-ity* and the native *-ness*, typically used in contemporary English for the creation of abstract nouns derived from adjectives (e.g. *curious* – *curiosity*; *happy* – *happiness*), and to a lesser extent also from other word categories, such as nouns (e.g. *authority* < Lat. *auctor* “author”, *witness*), pronouns (e.g. *quality* < Lat. *quālis* “of what kind”, *whoness*), verbs (e.g. *restority* < *restore*, *forgiveness*),² numerals (e.g. *oneness*, *unity*) and, in the case of *-ness*, also participles (e.g. *drunkenness*), adverbs (e.g. *backwardness*), phrases (e.g. *matter-of-factness*), and prepositions (*aboutness*). In spite of their apparent similitude, the two suffixes differ with regard to the kinds of bases they attach to (see, among others, Marchand, 1969: 314, 334–335; Aronoff, 1976: 36–38; Plag, 2003: 115–116), their semantics (Riddle, 1985; Romaine, 1985), and the kinds of registers in which they tend to appear (Baayen & Renouf, 1996; Cowie, 1998: 219–224; Plag et al., 1999; Gardner, 2014: 141–173), something which partly relates to the more learned and prestigious connotations of the borrowed form *-ity* and to the fact that it almost exclusively combines with Romance words, whose use is favoured in formal, written contexts (Biber et al., 1999: 325). Research also suggests that the use of the two suffixes may be related to sociolinguistic factors, with men displaying a consistent overuse of *-ity* compared to women throughout the history of English, whereas no such tendency is evident in the use of *-ness* (see Säily & Suomela, 2009; Säily, 2011, 2018).

Plenty of previous studies have explored the development of these two rival suffixes from both synchronic and diachronic perspectives. Notwithstanding the importance of register analysis in the development of languages (see, e.g. Biber & Gray, 2013), however, few have investigated register variation throughout the EModE period. Cowie’s (1998) analysis based on A Representative Corpus of Historical English Registers (see ARCHER, 1990–1993/2002/2007/2010/2013), for

1. For interchangeability among suffixes and the creation of similar doublets in ME, see Dalton-Puffer (1996: 126–130) and Gardner (2014: 29–31, 70–110, 263–271). Riddle (1985: 448–449) notes that, although interchangeability between *-ity* and *-ness* is not always possible, historically a number of minimal pairs existed with an apparently similar meaning.

2. Deverbal formations with *-ness* were relatively common in OE, although they lost currency after 1250 and are no longer productive nowadays (see Dalton-Puffer, 1996: 82–83; Gardner, 2014: 27 and references therein).

example, focused on data from 1650 onwards. Likewise, Palmer's (2009, 2015) insightful contributions cover writings produced across several registers between 1300 and 1600, yet no evidence is provided for the seventeenth century. In an attempt to fill this gap in the literature, Rodríguez-Puente (2020) used seventeen registers as a source of evidence to explore the development of the two suffixes during the EModE period, obtaining results which suggest that *-ity* gained ground on *-ness* between the sixteenth and eighteenth centuries. The change seems to have begun in formal written registers and spread towards speech-related ones, probably aided by a general trend towards the adoption of a more literate style particularly during the eighteenth century (Biber & Finegan, 1997; McIntosh, 1998: 23–24), which would arguably favour the use of the more learned and prestigious borrowed form *-ity*. However, this analysis was hampered by the lack of robust statistical methods for analysing variation in word-formation processes across subcorpora of different sizes. While Säily and Suomela (2009) introduced such a method for analysing variation within an individual word-formation process, no similar method has thus far been devised for comparing different processes.

In this paper we verify and elaborate on the results of Rodríguez-Puente (2020) by developing methods that enable the diachronic comparison of competing processes across subcorpora based on, for example, register or social category. The source corpora include the Corpus of English Dialogues (1560–1760; Kytö & Culpeper, 2006), the Penn-Helsinki Parsed Corpus of Early Modern English (1500–1710; Kroch et al., 2004), and the EModE section of the Corpus of Historical English Law Reports, 1535–1999 (Rodríguez-Puente et al., 2018), which cover a wide variety of formal and informal written and “oral” registers. As a general contribution to diachronic corpus research, we introduce a novel method of periodisation that enables the aggregation of data derived from historical corpora representing different but overlapping periods, thus facilitating comparability across datasets.

The paper is structured as follows. Section 2 includes a review of the relevant literature, focussing first on the historical development of the two suffixes, and then moving on to register analysis and morphological productivity as key issues for their study. Section 3 gives an account of the sources from which the examples examined were extracted. In Section 4 we present our methodology and the results obtained, first regarding usage differences along the speech-written continuum (4.2) and then along the formal-informal continuum (4.3). Finally, Section 5 provides a summary and some concluding remarks on the results and methods.

2. *-ity* vs. *-ness*: The effect of register on their morphological productivity

A large amount of previous research has helped to broaden our knowledge of the history of the competing suffixes *-ity* and *-ness*, based on both corpora (see, among others, Säily & Suomela, 2009; Säily, 2014, 2016, 2018; Rodríguez-Puente, 2020, and references therein) and dictionaries (see, among others, Riddle, 1985; Aronoff & Anshen, 1998; Lindsay & Aronoff, 2013, and references therein).

The native suffix *-ness* was already well established in OE, and by ME it was a very frequent and productive suffix (Dalton-Puffer, 1996: 128; Gardner, 2014: 71–76, 84–85, 113–115). It was, in fact, one of the first native suffixes to combine with Romance words, which probably contributed to its successful spread (Romaine, 1985): it can attach to both native and Romance bases, while examples of formations with Germanic bases and *-ity* are practically unattested.³ The Romance suffix *-ity*, for its part, entered the inventory of English suffixes during the early ME period, first adopted through French loans in *-(i)te*, and eventually words in *-te* were Latinised to *-ity* (Marchand, 1969: 312–314). Like other foreign derivational morphemes, *-ity* was predominantly used with foreign bases, and then progressively gained productivity in terms of tokens and types towards the end of the ME period (Dalton-Puffer, 1996: 106–107; Hundt & Gardner, 2017: 118–119), especially with adjectives in *-able*, and further during the course of the EModE period (Nevalainen, 1999: 398).

Previous work has shown that ever since *-ity* entered the inventory of deadjectival suffixes in ME times, a certain degree of competitiveness or “rivalry” has been at work between the two suffixes. The notion of such competition is supported by the attestation of pairs such as *vocalness* and *vocality* (see Marchand, 1969: 335).⁴ According to Riddle (1985), the emergence of doublets of this kind is due to the existence of a semantic distinction between the two suffixes: although the two suffixes were initially synonymous, their coexistence triggered a process of lexical diffusion which led *-ness* to acquire the meaning component “embodied trait” and *-ity* to indicate “abstract (or concrete) entity”. Cowie (1998: 259–261), however, finds no evidence of such diversification in EModE or even contemporary English. To her, the distinction between such pairs is one of register: both can describe a characteristic or attribute, but the *-ity* word is more appro-

3. Among the few exceptions, *betweenity*, *forlornity*, *oddity*, *queerity*, *scarcity*, and *womanity* are frequently mentioned in the literature (see, among others, Marchand, 1969; Dalton-Puffer, 1996: 106–107; Nevalainen, 1999: 398; Baeskow, 2012: 9–10; Gardner, 2014: 145). The earlier variant *-te* is also attested with native English bases, e.g. *evilty* (Gardner, 2014: 32).

4. For a comprehensive list of these pairs in ME, see Gardner (2014: 263–271).

priate for specialised terminology due to its learned character. She goes on to say that, although an entity meaning developed in words such as *rarity* but not *rareness*, the attribute meaning was not completely lost in the *-ity* words. According to Baayen (1993), *-ness* is more productive than *-ity* in contemporary English, especially in the written rather than the spoken medium (Plag et al., 1999: 224), although he also acknowledges that this may depend on the type of base (see also Baayen & Renouf, 1996; Plag, 2003: 116; Lindsay, 2012) and may respond to a complex interplay of phonological, morphological, semantic, and functional factors.⁵

2.1 Register variation

Following Cowie's (1998) register hypothesis to explain variation between *-ity* and *-ness*, we seek to explore the development of the two suffixes from the register perspective. The labels 'genre' and 'register' have been lately adopted to define different approaches to the analysis of texts (see especially Biber & Conrad, 2019). The genre approach focuses on the rhetorical structure of texts, whereas the register approach deals with the "functional relationships between the situations of use of a text variety and the patterns of language use" (Gray & Egbert, 2019: 1). Correspondingly, the term 'register' identifies "text varieties that are defined by the situational characteristics of a text and which, as a result, typically share similar linguistic profiles" (Gray & Egbert, 2019: 1–2). In other words, it refers to "a variety associated with a particular situation of use" (Biber & Conrad, 2019: 6) and implies a combined analysis of both written and oral productions based on context and the particular domain of discourse. Works on language variation and change have long assumed the existence of register differences, and a number of recent publications have effectively demonstrated that register can be a predictor of language change (see, among others, Biber, 2012; Biber & Egbert, 2016; Biber & Gray, 2013, 2016). Biber and Gray (2013: 104), for example, argue that registers are a mediating factor for diachronic change in language showing that "minor differences in register can correspond to meaningful and systematic differences in the patterns of linguistic change". As Kytö (2019: 155) rightly points out, registers have "been shown to act as powerful vehicles promoting or retarding language change, or even contributing to stability". For her, the relationship between register and linguistic evolution makes register a key concern for the historical approach to language.

A register-based approach to the development of *-ity* and *-ness* during the EModE period is thus crucial for the description of the two suffixes. The departing hypothesis of Rodríguez-Puente (2020) was that as a Romance, learned suffix,

5. See Säily (2014: 30–31) and references therein.

-ity would predominate in formal written registers, especially in those which are historically connected with a tradition of writing in Latin and French. However, the results of Rodríguez-Puente (2020) reveal a different picture. At the beginning of the seventeenth century, nominalisations in *-ness* dominate in all registers, except law reports and statutes, where *-ity* is the preferred form. There is, however, a clear change during the period: *-ness* derivatives lose ground progressively and are finally superseded by *-ity* formations in practically all registers towards the late seventeenth and early eighteenth centuries. As will soon become evident (Section 4), our results both verify and moderate the claims of Rodríguez-Puente (2020).

2.2 Morphological productivity

Morphological productivity was long ago defined by Bolinger (1948:18) as “the statistically determinable readiness with which an element enters into new combinations”. Token frequencies are typically used as a measure to trace the distribution and development of a language item diachronically and across registers, but they alone cannot be taken as a measure of productivity, since the overuse of a few individual formations can provide misleading results (Cowie & Dalton-Puffer, 2002: 414, 426). Type counts, on the other hand, constitute a reliable measure, since a productive suffix produces many different words or types (Dalton-Puffer, 1996:217). However, there is a correlation between the two measures: a suffix which is productive in terms of types is likely to occur more frequently (i.e. produce more tokens) than an unproductive one (Dalton-Puffer, 1996:217). ‘Hapax legomena’ or single-occurrence items can also be considered as a sign of the productivity of a construction but, as demonstrated by Baayen (1989, 1992), a measure of this kind is only appropriate for fairly large corpora. In small corpora, however, “productive and non-productive processes alike produce so few occurrences that there will be an overestimation of hapaxes, inflating the productivity counts” (Palmer, 2015: 109). A better way of calculating the productivity of a construction is by analysing the type/token ratio (TTR), although being a measure highly dependent on token counts it also has limitations (Baayen, 2008: 223). One of the problems stemming from using the TTR in diachronic corpora is that such an analysis does not take into account the new types that are introduced from one subperiod to the other. A more elaborate approach to measuring the productivity of derivatives over time was first developed by Cowie and Dalton-Puffer (2002) and successfully applied by Palmer (2015; see also Gardner, 2014), who employed the aggregation of new types to observe diachronic changes in type frequencies. Assuming that the data from the first corpus subperiod provides the “starting lexicon”, all the new types used for the first time in subsequent periods are counted.

The assumption is that if high rates of new types are added over a period of time, a suffix is likely to be productive over that period. New types in this analysis refer to neologisms in the corpus, though not necessarily neologisms in English. Given the size limitations of diachronic corpora, new types in this sense “serve a function similar to hapaxes in present-day studies of productivity” (Palmer, 2015: 115).

All type-based measures of productivity (types, hapaxes, TTR, new types) depend on corpus size in a non-linear manner (e.g. Baayen, 1992: 113; Säily, 2011: 127), which complicates comparisons between (sub)corpora of different sizes. This problem is especially pertinent to diachronic register or sociolinguistic studies, as they commonly deal with varying amounts of data from different periods, registers and/or social categories. One solution would be to downsample the larger subcorpora to match the size of the smallest subcorpus (Gaeta & Ricca, 2006), but this would make the already scarce data even scarcer. Another solution, employed by Plag et al. (1999), would be to use statistical modelling to inter- or extrapolate the growth curves of the processes analysed to make them comparable; however, models of this kind make the assumption that words occur randomly in texts, which is clearly incorrect and could lead to unreliable results. Moreover, there is no obvious way to estimate the statistical significance of the results (compare Kilgarriff, 2005).

To alleviate the aforementioned issues, Säily and Suomela (2009) introduced a computational method that builds growth curves for an individual word-formation process in a corpus, based on the non-parametric statistical technique of permutation testing. By building, say, a million curves for how the number of types (or another type-based measure) grows as we randomly sample more and more texts from the corpus, we obtain reliable estimates of “typical” numbers of types for each possible subcorpus size. We can then plot actual subcorpora onto these curves and see whether they fall within or outside the typical range, with a built-in measure of statistical significance: for instance, if a subcorpus has a greater number of types than 99.9% of the randomly composed subcorpora of the same size, its productivity is significantly higher at $p < 0.001$. Because the method samples entire texts rather than individual words, it preserves discourse structure and is free from the assumption that words occur randomly in texts. Nevertheless, the method has two major drawbacks: firstly, it only analyses variation *within* the productivity of an individual process (e.g. *-ity* suffixation), and secondly, it is not ideal for analysing diachronic change, since the *x*-axis of the growth curves represents corpus size rather than time, and the time periods plotted onto the curves cannot be compared with each other but only with the corpus as a whole. In the present paper, we therefore develop statistical and visual methods that better cater for the comparison of competing word-formation processes over time and across registers or social categories.

3. Sources

The examples used to measure the productivity of *-ity* and *-ness* in the present paper come from seventeen selected registers in three different corpora: the Corpus of English Dialogues (CED; 1560–1760; Kytö & Culpeper, 2006), the Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME; 1500–1710; Kroch et al., 2004) and the EModE section (1535–1749) of the Corpus of Historical English Law Reports, 1535–1999 (CHELAR; Rodríguez-Puente et al., 2018). Rodríguez-Puente (2020) selected these particular corpora on the basis that they contain samples which are representative of a wide variety of formal and informal registers distributed along the speech-written continuum. Although all the samples in the three corpora are recorded in written form, some are arguably closer to the spoken than to the written medium. Following Culpeper and Kytö's (2010: 17) classification, Rodríguez-Puente (2020) divided these registers into different sub-categories as a way to facilitate the analysis of the results. Conceived of within such a framework, Rodríguez-Puente (2020) represents these registers in Table 1.⁶

4. Methodology and analysis

To analyse the diachronic competition between *-ity* and *-ness* in speech-related vs. writing-based and writing-purposed registers, we introduce a new approach that addresses the challenges of comparing affixes across subcorpora with different amounts of data, and helps to establish the statistical significance of the findings. In brief, the essence of the new approach is this: we treat the competing suffixes as if they formed a linguistic variable and analyse the proportion of types representing the 'incoming variant', *-ity*, out of all *-ity* and *-ness* types. We show that this proportion grows nonlinearly with corpus size, which means that direct comparisons between registers are not justified. Therefore, for each time period, we compare the proportion of distinct *-ity* types out of all *-ity* and *-ness* types observed in one register category with the typical proportion of *-ity* types in a random subcorpus with the same total number of *-ity* and *-ness* types. To further analyse diachronic trends, we take samples of equal size from the register-based subcorpora and visualise the average proportion of *-ity* types over time.

For ease of explanation, we describe our statistical methods here while simultaneously presenting the results of our analysis of speech-related vs. writing-based and writing-purposed registers. We emphasise that the same methodology can

6. Note that the category "Law" comprises both statutes and law reports. For a full account of the process of selection of registers, see Rodríguez-Puente (2020).

Table 1. Distribution of registers in CED, PPCEME, and CHELAR according to the dimensions of (in)formality and their speech-like vs. written characterisation (from Rodríguez-Puente, 2020: 152)

		Informal	Formal
Speech-related	Speech-like	Diaries	
		Letters, private	
	Speech-based	Trial proceedings	
		Witness depositions	
	Speech-purposed	Drama	Sermons
Writing-based and writing-purposed			Bible
			Educational treatise
		(Auto)biography	History
			Law
		Travelogue	Letters, non-private
			Medicine
		Philosophy	
			Science

be applied in different contexts, and it does not assume e.g. a particular periodisation of the data, but for our case study we need to make some choices, and these are discussed in Section 4.1. The key new methodological ideas are described in detail in Section 4.2, together with the findings of our case study. Finally, Section 4.3 gives another application of the same methodology; there we analyse other hypotheses formulated by Rodríguez-Puente (2020) regarding the role of formal, particularly legal, registers in the rise of *-ity*.

4.1 Data collection and periodisation

The relevant examples were extracted from the corpora with *WordSmith Tools* (Scott, 2012) by searching for all the word forms containing one of the two endings in their different spellings, which were obtained from both wordlists based on the corpora themselves and the *Oxford English Dictionary* (OED). Following Säily and Suomela (2009: 90), we counted not only the words produced within

the English language by the addition of one of these suffixes but also those words that contained the suffixes etymologically and entered the language bearing them (e.g. ME *dignity* < Latin *dignitāt-em*). Each of the instances was associated with metadata on the corpus text in which it was found, including year, corpus period, word count, and register.

In order to increase the amount of data available for statistical analysis, we decided to combine CED and PPCEME, also adding CHELAR for the analysis of legal texts (Section 4.3 below). As the periodisations of the corpora did not match exactly, we used the year information and formed our own periods.⁷ To alleviate the issue of different start and end years of the corpora and other uncertainties in the years, we used a sliding window of 100 years that we moved at 25-year intervals in our analyses of change. Even though the sliding-window approach to periodisation has been used by e.g. Degaetano-Ortlieb and Teich (2018), we are not aware of previous work in historical corpus linguistics that would have used it to combine data from several corpora. The window size of 100 years was selected to ensure enough data per window, while the 25-year interval was chosen in order to gain a more fine-grained view of the phenomenon. The entire period analysed was 1500–1749, and within this period, our corpora included samples between the years 1560–1747 for CED, 1500–1719 for PPCEME, and 1544–1748 for CHELAR.

4.2 Speech-related vs. writing-based and writing-purposed registers

To analyse the competition between *-ness* and *-ity*, we started by calculating for each period the proportion of *-ity* types out of the sum of *-ness* and *-ity* types using the periodisation technique described above. This enabled us to quickly visualise the rise of *-ity* relative to *-ness* in speech-related registers, on the one hand, and in writing-based and writing-purposed registers, on the other, as illustrated in Figure 1. As discovered by Rodríguez-Puente (2020), speech-related registers seem to be lagging behind. However, there are two problems with this method: (i) we have different amounts of data from the two registers, so they are not directly comparable as the proportion of *-ity* could depend on corpus size in a non-linear manner (see Figure 2); and (ii) these calculations do not yet tell us whether the observed differences between the registers are statistically significant.

7. Sometimes the corpus year was uncertain and given as a range, e.g. 1592–1603; in such cases, we took the middle year of the range. For one year that was given as “ante 1671”, we looked for external information in the *Oxford Dictionary of National Biography* that enabled us to narrow it down to a range, of which we again took the middle year.

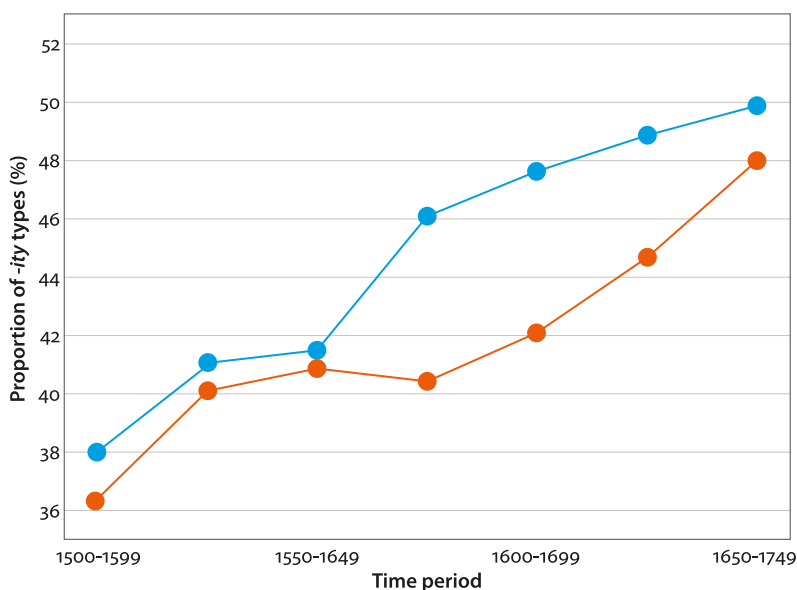


Figure 1. Proportion of *-ity* types out of *-ity* and *-ness* types in CED and PPCEME over time (orange = speech-related registers, blue = writing-based and writing-purposed registers); 100-year sliding window, 25-year intervals

In order to deal with these problems, let us first zoom in on an individual 100-year window, 1600–1699. For this corpus of texts written in subperiod 1600–1699, we can apply the method of Säily and Suomela (2009) to construct randomly sampled growth curves for the proportion of *-ity* types, again gaining reliable estimates of “typical” proportions for each possible subcorpus size, with which we can then compare actual subcorpora. Figure 2 visualises these curves along with the subcorpora of speech-related registers (orange dot) vs. writing-based and writing-purposed registers (blue dot). Note that the *x*-axis now represents corpus size (in terms of the number of *-ity* and *-ness* types) rather than time period. The shading indicates the areas in which 80% and 95% of random subcorpora fall; in other words, only 2.5% of the random subcorpora are above the shaded region and 2.5% below. We can see that the proportion of *-ity* types in the subcorpus of speech-related registers is very low in comparison with “typical” numbers in random subcorpora of the same size (orange dotted line), whereas the proportion in the subcorpus of writing-based and writing-purposed registers is quite high in comparison with “typical” numbers in random subcorpora of the same size (blue dotted line). In addition, the figure confirms our suspicion that like other type-based measures, the proportion of *-ity* types depends on corpus size in a non-linear manner; hence, we do need a more sophisticated method than that shown in Figure 1.

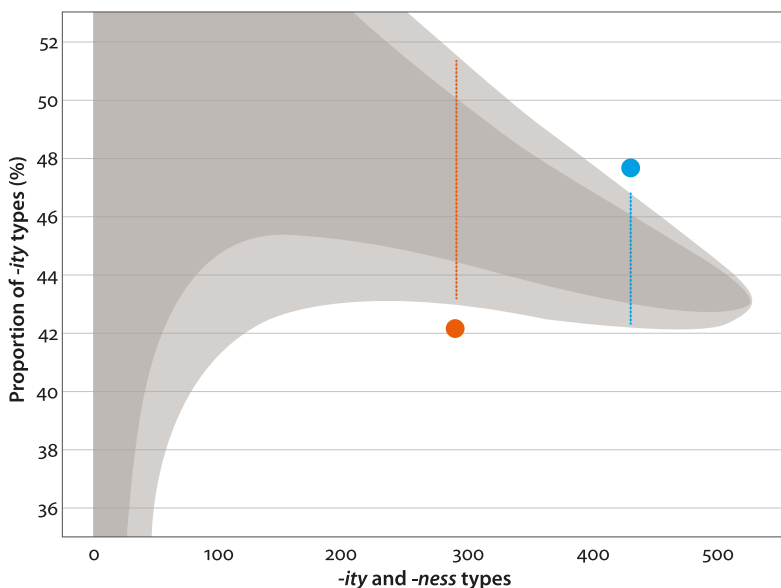


Figure 2. Randomly constructed growth curves for the proportion of *-ity* types in CED and PPCEME within the time period 1600–1699 (shaded regions = typical proportion in random subcorpora (the darker region covers 80% of the subcorpora and both regions together cover 95% of them); orange dot = speech-related registers; orange dotted line = typical proportion in random subcorpora of the same size; blue dot = writing-based and writing-purposed registers; blue dotted line = typical proportion in random subcorpora of the same size)

We have now solved problems (i) and (ii) for the time period 1600–1699: we have applied a method that accounts for subcorpus size, and we have discovered that the difference between registers is statistically significant. The next step is to do the same for all other time periods using our sliding window of 100 years and interval of 25 years. First, we consider speech-related registers, comparing for each period the observed proportion of *-ity* types in the subcorpus (orange dot in Figure 2) with the expected proportion in a subcorpus of that size (orange dotted line). We can then visualise diachronic change in both the observed and expected proportions, as in Figure 3 (the dotted line is only shown for 1600–1699 to make it easier to compare with Figure 2). It is easy to see that in speech-related registers the proportion of *-ity* is always on the low side, and it is significantly low around 1575–1699.

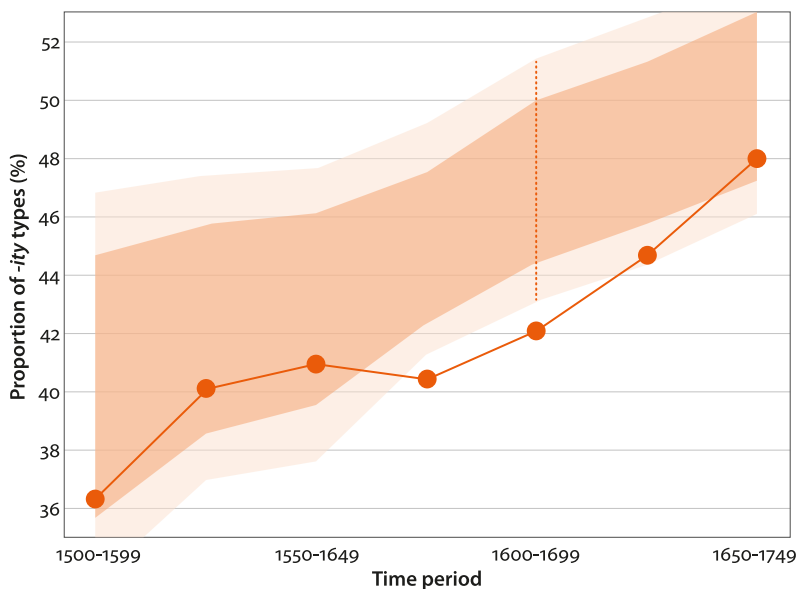


Figure 3. Proportion of *-ity* types out of *-ity* and *-ness* types in CED and PPCEME over time (solid line = speech-related registers; shaded regions = typical proportion in each period in random subcorpora of the same size; vertical dotted line = typical proportion in the time period 1600–1699)

In writing-based and writing-purposed registers, the opposite is the case. Figure 4 shows that the proportion of *-ity* is consistently high in this subcorpus, and significantly high around 1575–1699.

However, these figures do not enable a reliable assessment of trends over time: not only are the registers not directly comparable with each other, but neither are the time periods, since we have different amounts of data from each time period. We therefore add another step to our analysis and take samples of equal size from the register-based subcorpora. Figure 5 illustrates the average proportion of *-ity* types in both register categories over time, in samples of a total number of 100 *-ness* and *-ity* types. The figure shows that *-ity* does gain ground over *-ness* over time and that speech-based registers are lagging behind at first. The statistical significance of the lag was shown in Figure 3 for the period 1575–1699.

In order to gain a more fine-grained picture of the change, we can decrease our window size from 100 years to e.g. 50 years, as in Figure 6. To make sure all periods have data from both CED and PPCEME, we restrict our analysis to the years 1550–1725. It seems that the proportion of *-ity* starts a rapid rise in writing-based and writing-purposed registers in the period 1575–1625, so certainly by the

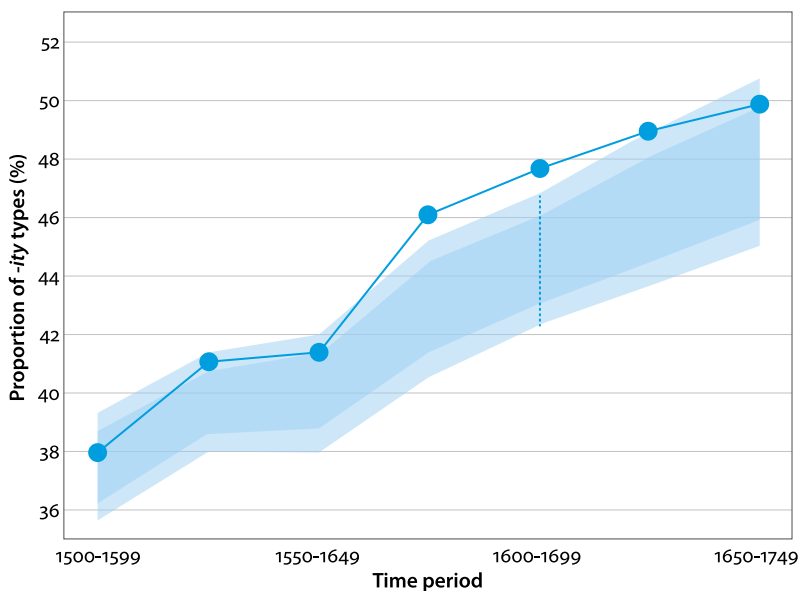


Figure 4. Proportion of *-ity* types out of *-ity* and *-ness* types in CED and PPCEME over time (solid line = writing-based and writing-purposed registers; shaded regions = typical proportion in each period in random subcorpora of the same size; vertical dotted line = typical proportion in the time period 1600–1699)

beginning of the seventeenth century, while speech-related registers lag behind until 1650–1699, or the end of the seventeenth century.

These results provide strong support to the findings by Rodríguez-Puente (2020) that *-ity* gained ground on *-ness*, and that the change began in written registers and spread towards speech-related ones. We now have more information on the timing of this development: while the proportion of *-ity* types seems to increase throughout the period examined, it starts a rapid rise in writing-based and writing-purposed registers by the beginning of the seventeenth century, whereas speech-related registers are lagging behind until the end of the seventeenth century. It should, however, be noted that although *-ity* gains ground, its proportion in either subcorpus does not exceed 50% even in the last 100-year period examined (1650–1749) when we consider the subcorpora in their entirety (Figures 3–4), so this measure does not fully support the claim by Rodríguez-Puente (2020) that *-ness* formations are eventually superseded by *-ity* formations. Moreover, before making claims about *-ity* becoming more productive than *-ness*, we should consider such factors as whether the formation was originally borrowed or derived within English, as well as different measures of productivity (see further Section 5 below).

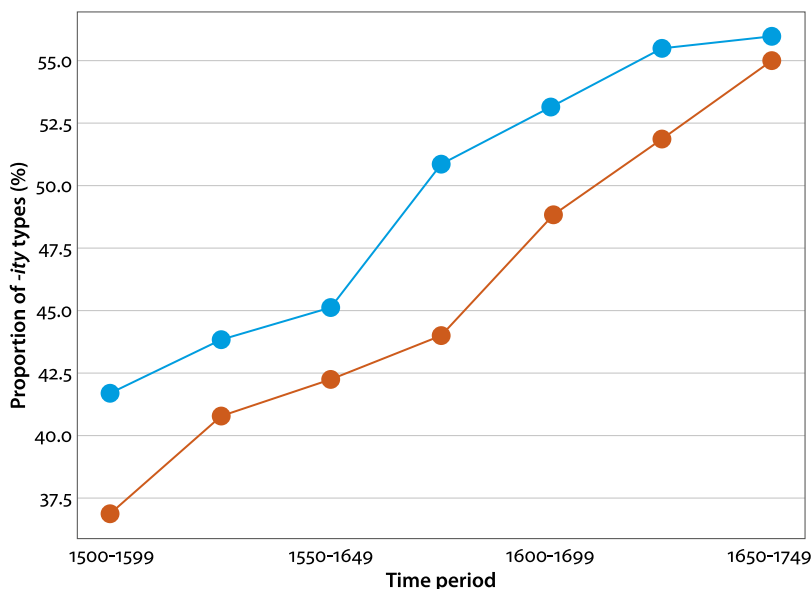


Figure 5. Proportion of *-ity* types out of *-ity* and *-ness* types in CED and PPCEME over time. Curves: Randomly sampled subcorpora with 100 distinct *-ity* and *-ness* types (orange = speech-related registers, blue = writing-based and writing-purposed registers); 100-year sliding window, 25-year intervals

4.3 Formal vs. informal registers and the role of legal texts

So far we have analysed the rise in the proportion of *-ity* in speech-related registers vs. writing-based and writing-purposed registers, but another relevant dimension is that of formal vs. informal registers (see Table 1). These dimensions are obviously related: most of the speech-related registers in our corpora can be characterised as informal, with only sermons on the formal side, whereas most of the writing-based and writing-purposed registers are formal, with only (auto)biography and travelogue belonging to a less formal category. Therefore, the results in Section 4.2 largely apply to formal vs. informal registers as well, so that informal registers are expected to lag behind, as also found by Rodríguez-Puente (2020). Delving deeper into this distinction, we attempted to analyse formality-based variation in the increase of *-ity* within speech-related registers, on the one hand, and within written registers, on the other. Here, however, the results of the statistical analysis were inconclusive owing to lack of data, as e.g. sermons only comprise 22 texts across the entire period covered by PPCEME.

One of the key findings of Rodríguez-Puente (2020) was that *-ity* seems to have begun its rise in the extremely formal register of statutes and law reports. To

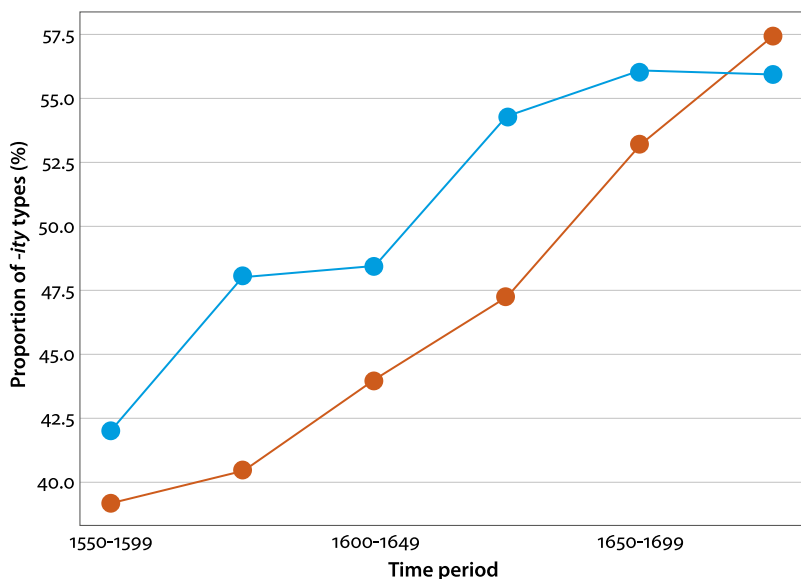


Figure 6. Proportion of *-ity* types out of *-ity* and *-ness* types in CED and PPCEME over time. Curves: Randomly sampled subcorpora with 100 distinct *-ity* and *-ness* types (orange = speech-related registers, blue = writing-based and writing-purposed registers); 50-year sliding window, 25-year intervals

test this hypothesis statistically, we added CHELAR to our repertoire of corpora and plotted the “Law” register against randomly composed subcorpora of the same size sampled from all three corpora (Figure 7). It is clear that the proportion of *-ity* in this register is significantly high throughout the period examined. The gap between the legal register and the random subcorpora is at its greatest towards the beginning of the period, indicating that the proportion of *-ity* was already quite high in legal texts before it started to increase in other registers. This supports the result obtained by Rodríguez-Puente (2020). However, as Rodríguez-Puente (2020) herself acknowledges, the potential influence of the legal register on this development must be taken with a pinch of salt. The high proportion of *-ity* in legal texts may reflect a preference for Romance vocabulary in this particular register since ME, whereas the rise of *-ity* in other registers could be connected to a general change towards a more elaborate and polished style during the eighteenth century (Biber & Finegan, 1997; McIntosh, 1998: 23–24).

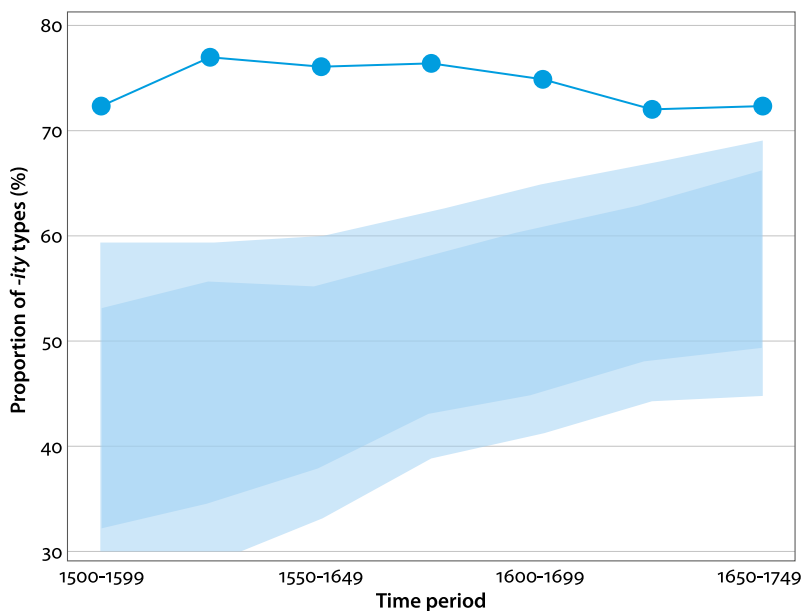


Figure 7. Proportion of *-ity* types out of *-ity* and *-ness* types in CED, PPCEME and CHELAR over time (solid line = “Law” register, i.e. statutes from PPCEME and law reports from CHELAR; shaded regions = typical proportion in each period in random subcorpora of the same size)

5. Discussion and conclusions

In this paper we have provided a new contribution to robust statistical and visual methods for the diachronic comparison of competing word-formation processes across registers. We have thus achieved our initial goal, that is, to develop a methodology which, on the one hand, could facilitate comparisons between corpora representing different but overlapping time periods, and, on the other, account for both subcorpus size and statistical significance. Applying this new methodology, we have been able to confirm the hypotheses put forward by Rodríguez-Puente (2020) that *-ity* gained ground on *-ness*, and that this occurred first in formal, written registers, and then spread to informal, speech-related ones. Our statistical analysis suggests that *-ity* took off in written registers by the beginning of the seventeenth century, while speech-related registers were lagging behind until the end of the seventeenth century. We have also confirmed the statistical significance of the high proportion of *-ity* in the register of extremely formal legal texts throughout the period examined. We can conclude, then, that our initial hypothesis that register might be a key factor in the development of the two suffixes during the EModE period is correct.

It must be noted, however, that the comparison between the two suffixes has not been made on a lexeme-by-lexeme basis, which would be complicated due to the relatively small size of the corpora analysed. Perhaps a more thorough analysis would need to involve exclusively competing forms created from the same base. However, the spread and increase of *-ity* derivatives in speech-related registers towards the eighteenth century could be a reflection of the development of a more literate and polished style, as part of the standardisation process which culminated in the prescriptivist period. Since *-ity* almost exclusively combines with Romance bases, the increase and spread of this suffix towards informal, speech-related registers may reflect a preference for Romance vocabulary, which would be more appropriate for a literate style.

Moreover, we have used the proportion of *-ity* types out of the sum of *-ness* and *-ity* types as the basis of our measure of productivity. As discussed in Section 2.1 above, *-ness* and *-ity* do not form a perfect linguistic variable in that they would always be interchangeable (see also Säily, 2014: 33–34). However, in this case where we are explicitly interested in how *-ity* gained ground on *-ness*, we argue that using the proportion of *-ity*, as if it was the proportion of an incoming variant out of a variable, is a justifiable measure. This argument is supported by the fact that we were able to gain such striking results which, being both consistent and for some periods highly statistically significant, are unlikely to be mere chance fluctuations.

Another issue is whether the proportion of all *-ity* types out of all *-ity* and *-ness* types is the best measure of productivity. In Section 2.2 we argued for the measure of new types rather than all types in the diachronic analysis of productivity. Rodríguez-Puente (2020) also used this measure, which could in part explain the discrepancy between her finding that *-ity* superseded *-ness* and our results, which show that both were used roughly equally by the end of the period examined. While the measure of new types may be better than that of hapax legomena in small historical corpora, its use is statistically questionable: which types occur in the starting lexicon may be largely a matter of chance, as may the number of new types per period, which in any case is liable to be so low that any statistical comparisons would remain inconclusive. The analysis of new types was also used by Cowie (1998: 219–224) to compare the development across registers of the two suffixes from the mid-seventeenth century to the end of the twentieth. Her results do not show any clear pattern of diachronic development, except for a remarkably high occurrence of new types of *-ity* derivatives in medical and scientific writing, which she attributes to the shift from an involved to an informational style, already identified by Biber and Finegan (1997). Cowie (1998: 221) also acknowledges that her analysis considers new types within registers, but that these “new types common to more than one register are less likely to be new in the language”.

In larger corpora, it would be of interest to apply our method using the proportion of new types (see Berg, 2021).

Although our methods provide strong statistical evidence to confirm Rodríguez-Puente's (2020) hypotheses, the limited amount of data in the corpora has precluded a more fine-grained register analysis. In future research, we may try to apply these methods to a larger dataset, such as *Early English Books Online* or *Eighteenth Century Collections Online*, although the sheer mass of material in databases of this kind might prevent a thorough examination of the search results, leading to an increased amount of noise in the data. It would also be of interest to trace the late history of the suffixes from the eighteenth century onwards to investigate whether the peak in *-ity* indeed corresponds to the shift towards a literate style in the eighteenth century.

To conclude, we wish to point out that the methods we have developed are applicable to diachronic corpus research beyond the issue of register variation in competing suffixes. Firstly, the statistical and visual methods presented in Section 4.2 above can be applied to any competing processes analysable via type-based measures, and the categories compared need not be based on register but can represent e.g. social or intra-linguistic factors instead – we have already conducted a small pilot study in which we have explored gender variation from a complementary perspective to Säily and Suomela (2009) and Säily (2016). Secondly, the method of periodisation introduced in Section 4.1 will facilitate the work of any scholar wishing to aggregate data from multiple corpora representing overlapping time periods and/or to maximise their data to more reliably analyse diachronic developments. The window size and interval should be determined on a case-by-case basis depending on the amount of data, the overlap between the corpora and the desired granularity of the results. In order to make these methods easier for the research community to adopt, and to make our results reproducible, the software and data we have used to conduct our analyses are freely available on *Zenodo* (Suomela, 2022a, 2022b; Rodríguez-Puente et al., 2022).

Funding

For generous financial support, we are grateful to the Spanish Ministry of Science and Innovation (grant PID2020-114604GB-100). This work was supported in part by the Academy of Finland (grant 323390).

Acknowledgements

We would like to thank the two anonymous reviewers and the editors of this issue for helpful feedback. Thanks are also due to participants at the ISLE 6 and ICAME 42 conferences for stimulating comments and discussions.

References


- ARCHER. (1990–1993/2002/2007/2010/2013). *A Representative Corpus of Historical English Registers*. Originally compiled under the supervision of Douglas Biber and Edward Finegan (Northern Arizona University and University of Southern California). Modified and expanded by members of a consortium of universities. Current consortium members: Universities of Bamberg, Freiburg, Heidelberg, Helsinki, Lancaster, Leicester, Manchester, Michigan, Northern Arizona, Santiago de Compostela, Southern California, Trier, Uppsala, and Zurich. <https://www.projects.alc.manchester.ac.uk/archer/>
- Aronoff, M. (1976). *Word Formation in Generative Grammar*. The MIT Press.
- Aronoff, M., & Anshen, F. (1998). Morphology and the lexicon: Lexicalization and productivity. In A. Spencer & A. M. Zwicky (Eds.), *The Handbook of Morphology* (pp. 237–247). Blackwell.
- Baayen, R. H. (1989). *A Corpus-based Approach to Morphological Productivity. Statistical Analysis and Psycho-linguistic Interpretation* [Unpublished doctoral dissertation]. Free University of Amsterdam.
- Baayen, R. H. (1992). Quantitative aspects of morphological productivity. In G. Booij & J. Van Marle (Eds.), *Yearbook of Morphology 1991* (pp. 109–149). Kluwer Academic Publishers. https://doi.org/10.1007/978-94-011-2516-1_8
- Baayen, R. H. (1993). On frequency, transparency and productivity. In G. Booij & J. Van Marle (Eds.), *Yearbook of Morphology 1992* (pp. 181–208). Kluwer Academic Publishers. https://doi.org/10.1007/978-94-017-3710-4_7
- Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511801686>
- Baayen, R. H., & Renouf, A. (1996). Chronicling the *Times*: Productive lexical innovations in an English newspaper. *Language*, 72(1), 69–96. <https://doi.org/10.2307/416794>
- Baekow, H. (2012). *-Ness and -ity*: Phonological exponents of *n* or meaningful nominalizers of different adjectival domains? *Journal of English Linguistics*, 40(1), 6–40. <https://doi.org/10.1177/0075424211405156>
- Berg, K. (2021). Productivity, vocabulary size, and new words. A response to Säily (2016). *Corpus Linguistics and Linguistic Theory*, 17(1), 177–187. <https://doi.org/10.1515/cllt-2017-0075>
- Biber, D. (2012). Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory*, 8(1), 9–37. <https://doi.org/10.1515/cllt-2012-0002>
- Biber, D., & Conrad, S. (2019). *Register, Genre, and Style* (2nd ed.). Cambridge University Press. <https://doi.org/10.1017/9781108686136>
- Biber, D., & Egbert, J. (2016). Register variation on the searchable web: A multi-dimensional analysis. *Journal of English Linguistics*, 44(2), 95–137. <https://doi.org/10.1177/0075424216628955>
- Biber, D., & Finegan, E. (1997). Diachronic relations among speech-based and written registers in English. In T. Nevalainen & L. Kahlas-Tarkka (Eds.), *To Explain the Present: Studies in the Changing English Language in Honour of Matti Rissanen* (Mémoires de la Société Néophilologique de Helsinki LII, pp. 253–275). Société Néophilologique.
- Biber, D., & Gray, B. (2013). Being specific about historical change: The influence of sub-register. *Journal of English Linguistics*, 41(2), 104–134. <https://doi.org/10.1177/0075424212472509>

- Biber, D., & Gray, B. (2016). *Grammatical Complexity in Academic English: Linguistic Change in Writing*. Cambridge University Press.
- Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Pearson Education.
- Bolinger, D. L. (1948). On defining the morpheme. *Word*, 4, 18–23.
<https://doi.org/10.1080/00437956.1948.11659323>
- Cowie, C. (1998). *Diachronic Word-formation: A Corpus-based Study of Derived Nominalizations in the History of English* [Unpublished doctoral dissertation]. University of Cambridge.
- Cowie, C., & Dalton-Puffer, C. (2002). Diachronic word-formation and studying changes in productivity over time: Theoretical and methodological considerations. In J. E. Díaz Vera (Ed.), *A Changing World of Words: Studies in English Historical Lexicography, Lexicology and Semantics* (pp. 410–437). Rodopi.
- Culpeper, J., & Kytö, M. (2010). *Early Modern English Dialogues: Spoken Interaction as Writing*. Cambridge University Press.
- Dalton-Puffer, C. (1996). *The French Influence on Middle English Morphology: A Corpus-based Study of Derivation*. Mouton de Gruyter. <https://doi.org/10.1515/9783110822113>
- Degaetano-Ortlieb, S., & Teich, E. (2018). Using relative entropy for detection and analysis of periods of diachronic linguistic change. In B. Alex, S. Degaetano-Ortlieb, A. Feldman, A. Kazantseva, N. Reiter, & S. Szpakowicz (Eds.), *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL-2018)* (ACL Anthology W18–45, pp. 22–33). Association for Computational Linguistics. <https://aclanthology.org/W18-4503/>
- Gaeta, L., & Ricca, D. (2006). Productivity in Italian word formation: A variable-corpus approach. *Linguistics*, 44(1), 57–89. <https://doi.org/10.1515/LING.2006.003>
- Gardner, A.-C. (2014). *Derivation in Middle English: Regional and Text Type Variation* (Mémoires de la Société Néophilologique de Helsinki XCII). Société Néophilologique.
- Gray, B., & Egbert, J. (2019). Register and register variation. *Register Studies*, 1(1), 1–9.
<https://doi.org/10.1075/rs.00001.edi>
- Hundt, M., & Gardner, A.-C. (2017). Corpus-based approaches: Watching English change. In L. J. Brinton (Ed.), *English Historical Linguistics: Approaches and Perspectives* (pp. 96–130). Cambridge University Press. <https://doi.org/10.1017/9781316286562.005>
- Kilgarraff, A. (2005). Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory*, 1(2), 263–275. <https://doi.org/10.1515/clt.2005.1.2.263>
- Kroch, A., Santorini, B., & Delfs, L. (2004). *Penn-Helsinki Parsed Corpus of Early Modern English* (PPCEME; 1st ed., release 3). Department of Linguistics, University of Pennsylvania. <http://www.ling.upenn.edu/ppche/ppche-release-2016/PPCEME-RELEASE-3>
- Kytö, M. (2019). Register in historical linguistics. *Register Studies*, 1(1), 136–167.
<https://doi.org/10.1075/rs.18011.kyt>
- Kytö, M., & Culpeper, J. (2006). *A Corpus of English Dialogues 1560–1760*. <https://www.engelska.uu.se/research/english-language/electronic-resources/english-dialogues/>
- Lindsay, M. (2012). Rival suffixes: Synonymy, competition, and the emergence of productivity. In A. Ralli, G. Booij, S. Scalise, & A. Karasimos (Eds.), *Morphology and the Architecture of Grammar: On-line Proceedings of the 8th Mediterranean Morphology Meeting (MMM8)* (pp. 192–203). University of Patras.


- Lindsay, M., & Aronoff, M. (2013). Natural selection in self-organizing morphological systems. In N. Hathout, F. Montermini, & J. Tseng (Eds.), *Morphology in Toulouse: Selected Proceedings of Décembrettes 7* (pp. 133–153). Lincom.
- Marchand, H. (1969). *The Categories and Types of Present-day English Word-formation* (2nd ed.). C.H. Beck. (Original work published 1960)
- McIntosh, C. (1998). *The Evolution of English Prose 1700–1900: Style, Politeness, and Print Culture*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511582790>
- Nevalainen, T. (1999). Early Modern English lexis and semantics. In R. Lass (Ed.), *The Cambridge History of the English Language, III: 1476–1776* (pp. 332–458). Cambridge University Press.
- Oxford Dictionary of National Biography*. Oxford University Press. <https://www.oxforddnb.com>
- Oxford English Dictionary*. OED Online. Oxford University Press. <https://www.oed.com>
- Palmer, C. C. (2009). *Borrowings, Derivational Morphology, and Perceived Productivity in English, 1300–1600* [Unpublished doctoral dissertation]. The University of Michigan.
- Palmer, C. C. (2015). Measuring productivity diachronically: Nominal suffixes in English letters, 1400–1600. *English Language and Linguistics*, 19(1), 107–129. <https://doi.org/10.1017/S1360674314000264>
- Plag, I. (2003). *Word-formation in English*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511841323>
- Plag, I., Dalton-Puffer, C., & Baayen, H. (1999). Morphological productivity across speech and writing. *English Language and Linguistics*, 3(2), 209–228. <https://doi.org/10.1017/S1360674399000222>
- Riddle, E. M. (1985). A historical perspective on the productivity of the suffixes *-ness* and *-ity*. In J. Fisiak (Ed.), *Historical Semantics, Historical Word-formation* (pp. 435–461). Mouton de Gruyter. <https://doi.org/10.1515/9783110850178.435>
- Rodríguez-Puente, P. (2020). Register variation in word-formation processes: The development of *-ity* and *-ness* in Early Modern English. *International Journal of English Studies*, 20(2), 147–169. <https://doi.org/10.6018/ijes.364261>
- Rodríguez-Puente, P., Fanego, T., López-Couso, M. J., Méndez-Naya, B., Núñez-Pertejo, P., Blanco-García, C., & Tamaredo, I. (2018). *Corpus of Historical English Law Reports 1535–1999* (CHELAR; version 2). Research Unit for Variation, Linguistic Change and Grammaticalization, University of Santiago de Compostela.
- Rodríguez-Puente, P., Säily, T., & Suomela, J. (2022). *Data for the article “New methods for analysing historical suffix competition across registers”* (Version 1.0.0) [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.5898202>
- Romaine, S. (1985). Variability in word formation patterns and productivity in the history of English. In J. Fisiak (Ed.), *Papers from the 6th International Conference on Historical Linguistics* (pp. 451–465). John Benjamins.
- Säily, T. (2011). Variation in morphological productivity in the BNC: Sociolinguistic and methodological considerations. *Corpus Linguistics and Linguistic Theory*, 7(1), 119–141. <https://doi.org/10.1515/cllt.2011.006>
- Säily, T. (2014). *Sociolinguistic Variation in English Derivational Productivity: Studies and Methods in Diachronic Corpus Linguistics*. Société Néophilologique.
- Säily, T. (2016). Sociolinguistic variation in morphological productivity in eighteenth-century English. *Corpus Linguistics and Linguistic Theory*, 12(1), 129–151. <https://doi.org/10.1515/cllt-2015-0064>


- Säily, T. (2018). Change or variation? Productivity of the suffixes *-ness* and *-ity*. In T. Nevalainen, M. Palander-Collin, & T. Säily (Eds.), *Patterns of Change in 18th-century English: A Sociolinguistic Approach* (pp. 197–218). John Benjamins. <https://doi.org/10.1075/ahs.8.12sai>
- Säily, T., & Suomela, J. (2009). Comparing type counts: The case of women, men and *-ity* in early English letters. In A. Renouff & A. Kehoe (Eds.), *Corpus Linguistics: Refinements and Reassessments* (pp. 87–109). Rodopi.
- Scott, M. (2012). *WordSmith Tools* (Version 6) [Computer software]. Lexical Analysis Software.
- Suomela, J. (2022a). *Code for the article “New methods for analysing diachronic suffix competition across registers”* (Version 1.0.0). Zenodo. <https://doi.org/10.5281/zenodo.5898974>
- Suomela, J. (2022b). *TypeRatio: Comparing competing suffixes* (Version 1.0.0) [Computer software]. Zenodo. <https://doi.org/10.5281/zenodo.5898976>

Address for correspondence

Paula Rodríguez-Puente
 Facultad de Filosofía y Letras
 University of Oviedo
 C/ Amparo Pedregal s/n
 Campus El Milán
 33011 Oviedo
 Spain
 rodriguezppaula@uniovi.es
 <https://orcid.org/0000-0002-7177-5984>

Co-author information

Tanja Säily
 Faculty of Arts
 University of Helsinki
 tanja.saily@helsinki.fi
 <https://orcid.org/0000-0003-4407-8929>

Jukka Suomela
 Department of Computer Science
 Aalto University
 jukka.suomela@aalto.fi
 <https://orcid.org/0000-0001-6117-8089>

Publication history

Date received: 29 September 2020

Date accepted: 25 January 2022

Published online: 19 August 2022