Comparing type counts:
The case of women, men and *-ity* in early English letters

Tanja Säily[1] *and Jukka Suomela*[2]

[1] Research Unit for Variation, Contacts and Change in English (VARIENG)
Department of English, University of Helsinki
[2] Helsinki Institute for Information Technology HIIT
Department of Computer Science, University of Helsinki

Abstract

This work is a case study of applying nonparametric statistical methods to corpus data. We show how to use ideas from permutation testing to answer linguistic questions related to morphological productivity and type richness. In particular, we study the use of the suffixes -ity *and* -ness *in the 17th-century part of the* Corpus of Early English Correspondence *within the framework of historical sociolinguistics. Our hypothesis is that the productivity of* -ity*, as measured by type counts, is significantly low in letters written by women. To test such hypotheses, and to facilitate exploratory data analysis, we take the approach of computing accumulation curves for types and hapax legomena. We have developed an open source computer program which uses Monte Carlo sampling to compute the upper and lower bounds of these curves for one or more levels of statistical significance. By comparing the type accumulation from women's letters with the bounds, we are able to confirm our hypothesis.*

1. Introduction

The linguistic case we study is as follows. We have two roughly synonymous suffixes, *-ness* and *-ity*, which are typically used for forming abstract nouns from adjectives, as in example (1) below.

(1) a. *generous* [ˈdʒɛnərəs] + *-ness* → *generousness* [ˈdʒɛnərəsnɪs]
b. *generous* [ˈdʒɛnərəs] + *-ity* → *generosity* [dʒɛnəˈrɒsɪtɪ]

The first suffix, *-ness*, is etymologically native, while *-ity* entered the language as a result of contact with French during the Middle English period, and was later reinforced by loans from Latin (Marchand 1969: 312–313). The foreignness of *-ity* can be readily discerned from the above example: it changes the form of its base from [ˈdʒɛnərəs] to [dʒɛnəˈrɒs], whereas with *-ness* there is no change (but see Section 2.1 below). In addition, the meaning of words in *-ity* is often not entirely compositional, i.e., not deductible from the meanings of the base and the suffix. Thus, it is both (morpho)phonologically and semantically more opaque than *-ness* (cf. Riddle 1985: 443–444; Aronoff and Anshen 1998: 246).

What we are interested in doing with the suffixes is to compare their morphological productivity, a concept famously defined by Bolinger (1948: 18) as "the statistically determinable readiness with which an element enters into new combinations". More specifically, we wish to examine whether the productivity of each suffix varies between different sociolinguistic groups, as defined by Labovian sociolinguistic categories such as age, gender and social status. Many linguistic features show sociolinguistic variation, but to date this has been studied little in the case of morphological productivity, and not at all with the otherwise closely scrutinised pair of *-ness* and *-ity*.

Our data come from the 17th-century part of the *Corpus of Early English Correspondence* (1998; henceforth known as the *CEEC*). We have chosen personal letters as our material because they are one of the closest genres to speech, which is the primary medium of language and the most fertile ground for linguistic change (Nevalainen and Raumolin-Brunberg 2003: 28). This time period is interesting because it is to be expected that *-ity* would by this time have spread to wider use from the more literary genres in which it entered the language. Furthermore, a pilot study by Säily (2005) using the smaller *Corpus of Early English Correspondence Sampler* (1998) showed a gender difference in the use of *-ity* in letters of the 17th century.

We believe that *-ity*, as a learned and etymologically foreign suffix, is less productive with poorly educated social groups, such as women and the lower ranks, than with well-educated groups, such as men and the higher ranks. As to the productivity of *-ness*, we do not expect to find significant differences between social groups.

1.1 Objectives

The main measure of morphological productivity used in this study is that of type counts, i.e., how many different words in *-ity* and *-ness* are used by the different social groups. We seek to study the productivity of the suffixes *-ity* and *-ness* in our material by two complementary means:

1. Statistical hypothesis testing. We aim to formulate and test a hypothesis which captures our belief that gender is significant in the case of *-ity*.

2. Exploratory data analysis. Regardless of whether gender proves to be significant or not, we are interested in studying the correlation between productivity and a number of other variables, such as the age, domicile or social rank of the writers.

We present a unified approach which enables us to tackle both of these tasks.

1.2 Contributions

This work is a case study of applying nonparametric statistical methods to corpus data. We show how to use ideas from permutation testing to answer linguistic questions related to productivity and type richness. The basic techniques are standard but not widely used in the study of these questions – our hands-on report aims at promoting the use of these powerful tools. With this goal in mind, we have chosen to describe in detail one particular application of these techniques. The emphasis is on depth, not breadth: instead of side-tracking and discussing a number of alternative techniques at each point, we make particular choices and go through all the subtleties that need to be taken into account. We assume a basic knowledge of statistical hypothesis testing, but we have included an informal introduction to permutation tests.

We take the approach of computing accumulation curves for types and hapax legomena (i.e., types that occur only once). In particular, we use Monte Carlo sampling to compute the upper and lower bounds of these curves for some predetermined levels of statistical significance. Once we have computed an accumulation curve, we can test a hypothesis by simply plotting a data point on the curve. Exploratory data analysis is equally straightforward, and we can also qualitatively study the shape of the accumulation curves.

One of the main technical contributions is described in Section 5: we have developed a computer program which can be used to compute the curves. This is the only part of the method described here which is computationally intensive. In the implementation, the emphasis is on computational efficiency. The program is freely available under an open source licence.

The results achieved by using these methods on our data are reported in Section 6. As we shall see, we can conclude that our hypothesis is true: the type richness of *-ity* is indeed significantly low in the subcorpus which consists of women's texts. Exploratory data analysis reveals an unanticipated feature of the data: the type richness of *-ity* is also significantly low in the subcorpus which consists of the letters written in 1600–1639.

2. Background and related work

In this section, we justify the use of type counts for measuring morphological productivity, place the study in the framework of historical sociolinguistics, and review related work on using similar methods.

2.1 Type counts as a measure of morphological productivity

According to Dalton-Puffer (1996: 217), there is an obvious correlation between productivity and type counts: "a productive morphological rule produces many

different words (types), and it is therefore likely that in a given corpus a productive suffix will occur more often than an unproductive one". Type counts are by no means a perfect measure of productivity, however. As Cowie and Dalton-Puffer (2002: 416) point out, the existence of a large number of types may be due to aggregation through productivity in the past rather than current productivity. Furthermore, in the case of *-ity*, some words have been borrowed from French or Latin as a package including the suffix, with no productivity involved at all in English. This applies to the word *generosity* in our example (1): according to the *Oxford English Dictionary* (henceforth the *OED*), *generosity* has been in the language since about 1432 and is an adaptation of the Latin word *generōsitāt-em*.

Nevertheless, type counts are frequently used as a measure of productivity, for example by Baayen and Lieber, who call it the extent of use (1991: 818). This measure may not give us a full picture of the productivity of a suffix, but it can certainly be useful despite the above caveats about past productivity and borrowing. In addition, the impact of these caveats could be reduced by restricting the kinds of words that are counted. One possible restriction would be that the suffixed word must have had an extant base at the time when the material was written; another could be that the word must not have been in the language for, say, more than a century, as evidenced by its first attestation date in a major dictionary such as the *OED* (Cowie and Dalton-Puffer 2002: 419). These restrictions would increase the probability that the word in question was formed productively from suffix and base rather than retrieved as a whole from the mental lexicon of the writer.

For this study, however, we have elected to omit the above restrictions and count all words that etymologically contain the suffix in question – as noted by Plag (1999: 29), dropping out "non-productive formations" could mean prejudging the issue of whether the suffix is productive. The latter of the above restrictions at least would certainly be too limiting: To an individual user of the language, a word can be new even if it has been around in the language community for hundreds of years (cf. Baayen and Renouf 1996: 77), and thus even established words can be formed productively by users from the base and the affix. In fact, even if an affixed word exists in the mental lexicon of the user, he or she may still end up forming it from its constituents, depending on how frequent the affixed word is compared with its base – Hay (2001) claims this is true for processing (e.g., when reading), but we think it holds for producing words as well.

As for words with no extant base, they too may contribute to keeping the suffix productive, as they contain its form and meaning, and there is often an adjective related to the missing base that could be seen as the base by the user; see (2).[1] Various restrictions on type counts are explored in Säily (2008: 87–95).

(2) *ambiguity ~ ambiguous + -ity*

2.2 Historical sociolinguistics and morphology

The application of sociolinguistics to historical material is a fairly new approach: according to Nevalainen and Raumolin-Brunberg (2003: 2), the first systematic attempt at this was made by Suzanne Romaine in 1982. Nevalainen and Raumolin-Brunberg themselves are pioneers in this field, which is now called historical sociolinguistics. While morphology has been studied within this framework, research has so far concentrated on inflectional morphology such as the use of third-person *-s* vs. *-th* (Nevalainen and Raumolin-Brunberg 2003). Březina (2005) is a rare example of a study on the productivity of derivational prefixation from the perspective of historical sociolinguistics. To our knowledge, there have been no studies on suffixation from a similar perspective.

2.3 Methodology

Previous work on comparing the productivity of an affix between subcorpora often relies on the subcorpora being approximately the same size, so that for instance type counts obtained from each subcorpus can be compared directly. Then, if the type counts differ by an order of magnitude, it may be possible to draw conclusions without paying attention to statistical significance (e.g., Dalton-Puffer 1996: 106).

Empirically validated assumptions on modelling productivity have been made by, e.g., Baayen (1992, 1993). For example, the growth rate of the type accumulation curve has been approximated as the ratio between the number of hapax legomena and the total number of tokens with the affix (Baayen 1992: 115). Baayen (2001) studies both parametric and nonparametric models for the class of LNRE (large number of rare events) distributions, such as lexical frequency distributions. These models are based on the assumption that individual words appear randomly in texts; such modelling assumptions make it possible to extrapolate beyond observed sample size. For a recent study on the statistical models for the accumulation of types and hapax legomena, see Evert and Baroni (2005), and for related statistical software, see Evert and Baroni (2007).

Nonparametric methods similar to ours – in particular, Monte Carlo sampling of permutations – have been used in corpus linguistics to some extent. For example, Baayen (2001: 6–7, 24–32) computes Monte Carlo confidence intervals for the accumulation curves of some lexical characteristics. Permutations are generated at the level of individual words, which is consistent with the assumption that individual words appear randomly in texts. However, in many cases the observed values lie outside the confidence intervals (Baayen 2001: 6–7, 24–32; Tweedie and Baayen 1998: 335), indicating that the assumption of randomness causes bias in the results. Tweedie and Baayen (1998) address the bias by permuting words within a randomisation window. Our approach is to leave the original discourse structure intact and permute only large parts of the corpus.

Analogous research questions arise and similar methods can be used in studies of biodiversity in the field of ecology, to enable comparisons of species richness in different areas (see, e.g., Gotelli and Colwell 2001). Our text length corresponds to their number of individual animals; our number of types to their number of observed animal species; our two subcorpora of men and women to their different areas; and our type accumulation curves to their species accumulation curves.

3. Material

Our material in this study comes from the 17th-century part of the 2.7-million-word *Corpus of Early English Correspondence* (1998 version). The *CEEC* is an electronic collection of 6,039 letters composed by 778 writers between the years 1410?–1681. It was compiled by Terttu Nevalainen (team leader), Jukka Keränen, Minna Nevala (née Aunio), Arja Nurmi, Minna Palander-Collin and Helena Raumolin-Brunberg. Due to a lack of resources for transcribing and editing, the corpus is based on published editions of letters; however, some of the material has been checked against the originals by members of the *CEEC* team.

The *CEEC* is designed for studying the English language – more specifically, English English – in its socio-historical context. To this end, the writers have been carefully selected to give as balanced a representation of different social categories as possible. Nevertheless, the dominance of men from the upper ranks has been unavoidable: they were the most literate group, they were considered important enough that their letters were preserved, and their letters were later considered important enough to be published.

The 17th-century part of the *CEEC* consists of 1.4 million words covering the years 1600–1681. Unfortunately, only about a quarter of this material was written by women, as can be seen from Figure 1. The situation between different ranks, regions, etc. is similarly imbalanced.

Example (3), from a letter written in 1654 by Dorothy Osborne, illustrates the raw material in the corpus (emphases added).

(3) … to Visett a place you are soe much concern'd in, and to bee a *wittnesse* your selfe of the *probabillity* of your hopes though I will beleive you need noe other inducement to this Voyage then …
(A 1654 FN DOSBORNE 130:Heading)

For the purposes of this work, we have divided the corpus into **samples**, each consisting of one person's letters from a 20-year period in the corpus: 1600–1619, 1620–1639, 1640–1659, and 1660–1681. As an example, all letters in the corpus that were written by Dorothy Osborne in 1640–1659 form a sample called DOSBORNE-1640.

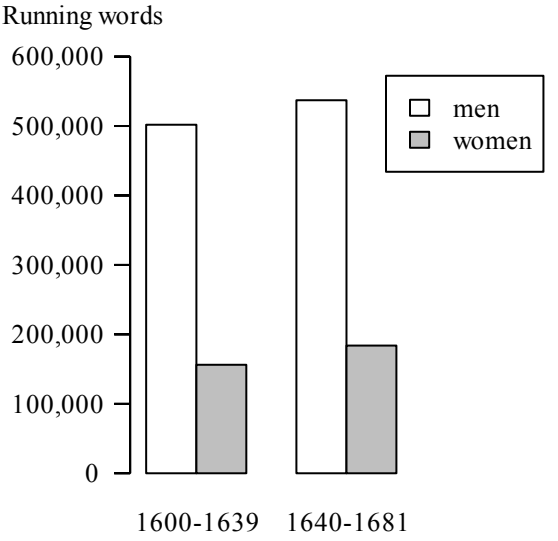

Figure 1. Running words written by men vs. women in the *CEEC*, 1600–1681.

3.1 Input data

The instances of *-ity* and *-ness* were extracted from the corpus using the WordCruncher program. Since the corpus was unlemmatised, and the grammatically tagged version was not yet available, this had to be done by searching for all word-forms which had a suitable ending. Different spelling variants of the suffixes were collected from the *OED*, the *Middle English Dictionary* (*MED*) and by browsing the corpus itself, after which they were used one by one in WordCruncher searches. Some of these variants, such as *-nes*, yielded a vast number of erroneous results, because many other words besides those having the suffix ended in that way, such as plurals of words ending in *-n*. These had to be weeded out by hand.

A combination of manual work and Perl scripts was used to produce a computer-readable list enumerating all instances of the suffixed words in a normalised form for each sample. The word *probabillity* in example (3) counts as one instance of the normalised form *probability* in the sample DOSBORNE-1640. There was a total of 94 occurrences of *-ity* in this sample, and they were instances of 31 different normalised forms, shown in example (4) below. Thus, we say that the number of *-ity* **tokens** is 94 and the number of *-ity* **types** is 31 for the sample DOSBORNE-1640.

(4) *antiquity authority calamity charity civility commodity conformity contrariety curiosity equality extremity formality gravity importunity impossibility infirmity insensibility necessity nobility opportunity piety possibility probability quality quantity reality severity society university vanity variety*

The information extracted from the corpus can be summarised as two incidence matrices, one for *-ity* and another for *-ness*. Each row of a matrix corresponds to one sample and each column corresponds to one type. The element at row *i* and column *j* indicates the number of occurrences of type *j* in sample *i*. The sum of the elements on row *i* equals the number of tokens in sample *i*, and the number of nonzero elements on row *i* equals the number of types in sample *i*. This is exemplified for *-ity* in Table 1.

Table 1. Part of the matrix representation of *-ity*.

	…	*contrariety*	*credulity*	*curiosity*	…	*probability*	…
…							
ASTUART-1600		0	0	1		0	
DOSBORNE-1640		1	0	4		1	
SPEPYS-1660		0	1	0		1	
…							

The number of running words was counted for each sample; for DOSBORNE-1640, the number of running words is 71,299 – the number of distinct words in the sample is not needed in our study. Sociolinguistic information on each person was retrieved from an auxiliary database; this included gender, domicile and social rank. For DOSBORNE-1640, the gender is 'female', the domicile is 'other' and the social rank is 'gentry upper'.

Our incidence matrices for *-ness* and *-ity* are freely available for download (Säily and Suomela 2007).

3.2 Characteristics of the input data

The total number of samples in the corpus is 412, of which 112 consist of letters written by women. The total number of different types of *-ity* in the corpus is 192 and the total number of different types of *-ness* is 312.

The relative sizes of the samples are illustrated in Figures 2 and 3. In the figures, samples from men are represented by white boxes, while samples from women are grey diamonds. The size of the symbol is in proportion to the number of running words in the sample. The largest samples are labelled, including DOSBORNE-1640 with 71,299 running words, and ASTUART-1600, Arabella Stuart's letters written in 1600–1619, with 30,472 running words.

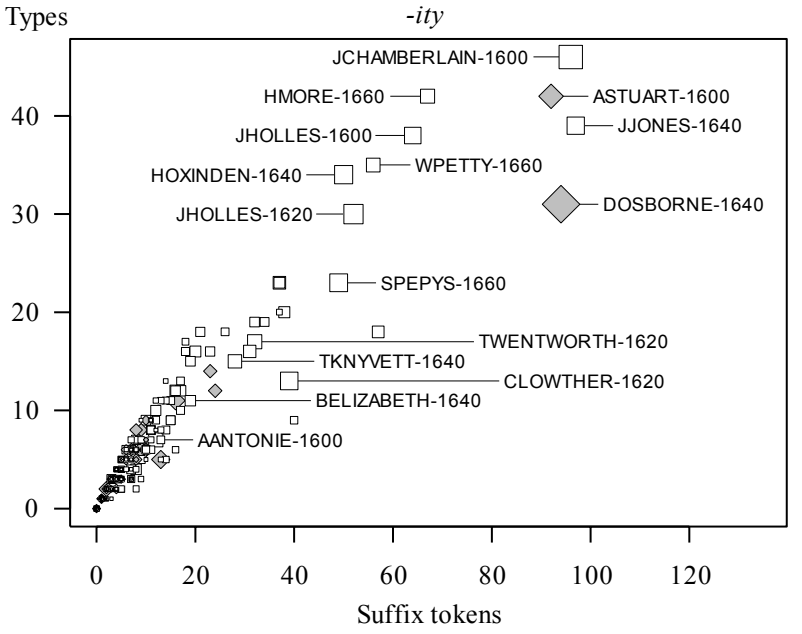

Figure 2. Samples ordered by the number of *-ity* types per *-ity* tokens.

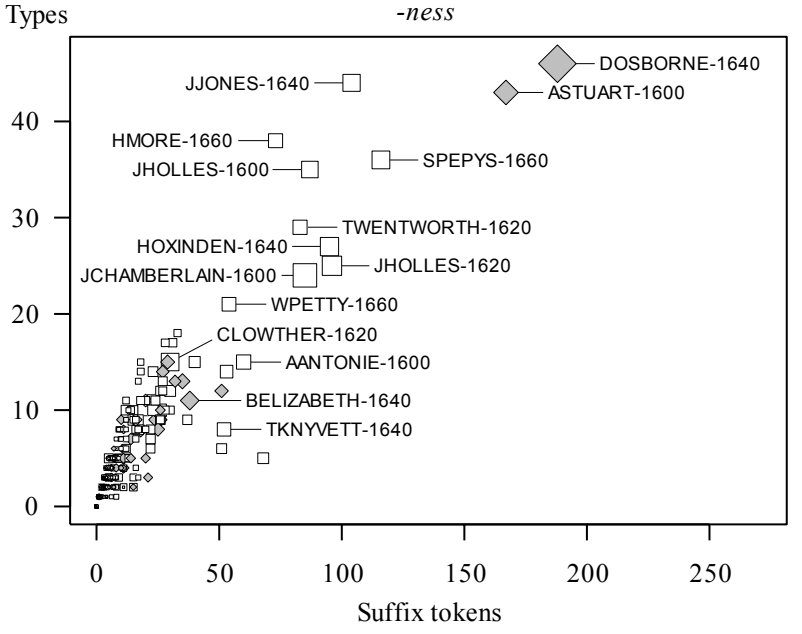

Figure 3. Samples ordered by the number of *-ness* types per *-ness* tokens.

Figure 2 presents the samples ordered by the number of *-ity* types they contain per *-ity* tokens. As noted above, there are 31 *-ity* types and 94 *-ity* tokens in the sample DOSBORNE-1640. Figure 3 presents the same information for *-ness* types. For example, there are 46 *-ness* types and 188 *-ness* tokens in DOSBORNE-1640. As can be seen from the figures, the size of the samples varies widely; there are many samples with very few tokens and types, and a few samples with very many tokens and types. From these figures we may observe, e.g., that while DOSBORNE-1640 includes more *-ness* types and tokens than any other sample, there are many samples from men which have a larger number of *-ity* types than this sample.

4. Methods

We are interested in comparing the productivity of a suffix between different subcorpora which consist of several samples, for example, all letters written by women. Our primary measure of productivity is the number of types. In the previous section we defined type counts for samples; this extends naturally to a whole subcorpus. As an alternative measure of productivity, we consider the number of hapax legomena. In precise terms, the measures are as follows; here we use the case of *-ity* as an example.

(a) Number of types. This is the number of different types of *-ity* which occur in the subcorpus at least once. For example, if the subcorpus contains occurrences of the word *generosity* (no matter how many times, regardless of the spelling) and no other *-ity* words, the number of types is 1.

(b) Number of hapax legomena or hapaxes. This is the number of different types of *-ity* which occur in the subcorpus exactly once. For example, if the subcorpus contains only one occurrence of the word *instability*, one occurrence of the word *capability*, four occurrences of the word *generosity* (in various spellings) and no other *-ity* words, the number of hapaxes is 2.

If we view the subcorpus as a matrix where the element at row i and column j indicates the number of occurrences of type j in sample i (recall Table 1), we can give the following equivalent definitions. Form a vector $\mathbf{v}$ by adding up all rows of the matrix. Then the number of types is the number of nonzero elements in $\mathbf{v}$, and the number of hapaxes is the number of elements equal to 1 in $\mathbf{v}$.

4.1 Comparing productivity between subcorpora

The measures we defined above have an obvious drawback: they are sensitive to the size of the subcorpus. In our material we have 80 types of *-ity* in the texts written by women and 183 types of *-ity* in the texts written by men; however, we cannot immediately say that the type richness of women's texts is lower, as we have much more material from men (see Figure 1 above).

Furthermore, the relation between the size of the subcorpus and the number of types occurring in it is not necessarily linear. Put simply, at the very beginning of the type accumulation curve, each *-ity* word is likely to be new, but later we are more likely to meet *-ity* words which have already occurred in the corpus. With hapaxes, the measure might even decrease as the size of the subcorpus increases. We shall see practical examples of the nonlinear behaviour throughout this work (e.g., Figures 4, 6 and 8 below).

Therefore, attempts to normalise the number of types by, say, dividing by the number of running words are not justifiable (cf. Gotelli and Colwell 2001). Indeed, such attempts give completely misleading results with our data. For example, the number of *-ity* types per 100,000 running words is approximately 23.5 for women and 17.6 for men in our material. It would appear that the type richness is higher for women, even though the opposite is the case, as we shall see.

We might be able to tackle the problem by making further modelling assumptions on the process which generates the text; we might, for example, assume that the occurrences of the words are independent, and we could then use the input data to estimate the probabilities of each person producing a particular word; this way we could compare the productivity of different persons. However, we are reluctant to make such simplifying assumptions, as the choice of words may have subtle dependencies on the textual context (see, e.g., Baayen 2001: 163).

We take a somewhat extreme approach in assuming nothing. Instead of trying to compare subcorpora of different sizes, we only assume that we can compare subcorpora of equal sizes. We use the following alternative definitions for equal size:

(i) The same number of running words.

(ii) The same number of *-ity* tokens.

For most of this work we focus on definition (i) in conjunction with measure (a), i.e., the number of types. Other combinations may also be of interest, and we can experiment with them by using the same general approach and the same tools. For example, if we use measure (b) and definition (ii), we compare the number of *-ity* hapaxes in subcorpora with the same number of *-ity* tokens. Equally well, we could compare the ratios between *-ity* hapaxes and *-ity* tokens, arriving at Baayen's (1992: 115) definition.

4.2 Statistical significance

We are not interested in merely noticing that a particular subcorpus has a lower number of types in comparison with another subcorpus. We are interested in differences which are statistically significant; informally, not likely to be mere random artefacts of the data.

We now review some basics of statistical hypothesis testing and apply the ideas to our problem. Let us choose the measure of productivity (a), the number of types, and say that we are willing to compare only subcorpora which are equal by definition (i), the number of running words. The idea that women are significantly less productive than men in this material is captured as follows. Let n be the number of running words in the subcorpus which consists of the texts written by women and let t be the number of types in this subcorpus.

> **Hypothesis.** Gender is significant. For a subcorpus with n running words, t is a particularly low number of types.

The null hypothesis is that there is no connection between the number of types and gender; the effect is caused by chance. More formally, the null hypothesis is that the numbers of running words and the rows of the incidence matrices for men and women are samples from the same population.

Intuitively, the null hypothesis suggests that the subcorpus of texts written by women could be constructed through the following process. We randomly pick samples from the corpus, labelling them as having been written by women, until the subcorpus we have accumulated is of size n; the rest of the corpus is then labelled as having been written by men. We emphasise that our samples consist of complete letters. We need not assume that the words within each letter are independent of the context; we only assume that samples as a whole are interchangeable under the null hypothesis.

We can test the hypothesis by estimating how likely it is that a subcorpus constructed in this way has as few as t types (we apply one-sided testing here). If this turns out to be very unlikely, say, happening on average only once in 100 trials, we reject the null hypothesis and accept the original hypothesis, with $p = 0.01$.

There is a subtlety: as we work at the granularity of samples, and the sizes of the samples vary, it may be that very few labellings – maybe just the original labelling – produce a subcorpus with exactly n running words. In practice, we make a minor adjustment. Informally, we consider subcorpora with at least n running words and not many more than that; making the subcorpus longer certainly cannot have a negative bias on the number of types. The case of hapaxes is more complicated; we come back to this issue in Section 5.3.

4.3 Permutation testing

Now, we have formalised our hypothesis and we are ready to do standard hypothesis testing – all we need to do is estimate the probability p of obtaining such an extreme case as at most t types in a subcorpus with n running words.

As we are dealing with type counts, we do not have a simple mathematical formula for calculating *p*: the probability depends not only on summary information such as the values *t* and *n* but on the full incidence matrix. Therefore, we use techniques from permutation testing (see, e.g., Good 2005). Applied to our problem in a straightforward manner, the basic idea would be as follows. We take the intuitive idea of picking samples in a random order quite literally. The order in which we pick the samples forms a permutation (reordering) of the samples. To calculate the probability *p*, we need to calculate the percentage of permutations which have at most *t* types in the first *n* running words. We generate all permutations of the samples, check which of them satisfy this condition, and compute the percentage *p*. The next section adapts this basic idea to our needs.

5. Implementation

Standard permutation testing would indeed suffice if all we were interested in was testing one hypothesis. However, we are also interested in exploratory data analysis. We want to consider several variables besides gender and see if they correlate with the number of types. Ideally, we would prefer to avoid repeating extensive computations between each experiment. We also wish to gain more understanding on the accumulation of types as a function of corpus size.

We can address all of these requirements by calculating type accumulation curves similar to that shown in Figure 4. This is the output generated by the computer program that we present in this section. First we describe how to interpret and use these curves; then we discuss the implementation which is used to compute the curves.

Figure 4 shows upper and lower bounds for the number of *-ity* types. On the *x* axis, we have the number of running words in the subcorpus. The bounds are plotted for various levels of statistical significance. For example, the solid black curve corresponds to the level $p = 0.01$; the lower bound for, say, 600,000 running words at this level is 123, and the upper bound at this level is 163. This can be interpreted as follows: in all permutations of the samples that we can construct from the whole corpus, less than 1% have fewer than 123 *-ity* types within the first 600,000 running words, and less than 1% have more than 163 *-ity* types within the first 600,000 running words. The *p* values here refer to a one-sided test; for a two-sided test, the *p* values need to be doubled.

Once we have computed the curves, we can immediately use them for hypothesis testing, in a very straightforward manner: we simply plot the data point that corresponds to the subcorpus of interest on these curves and see whether the point lies, for example, below the lower bound. If so, we conclude that the number of types is significantly low for a subcorpus of this size. This is merely an (indirect) application of a permutation test.

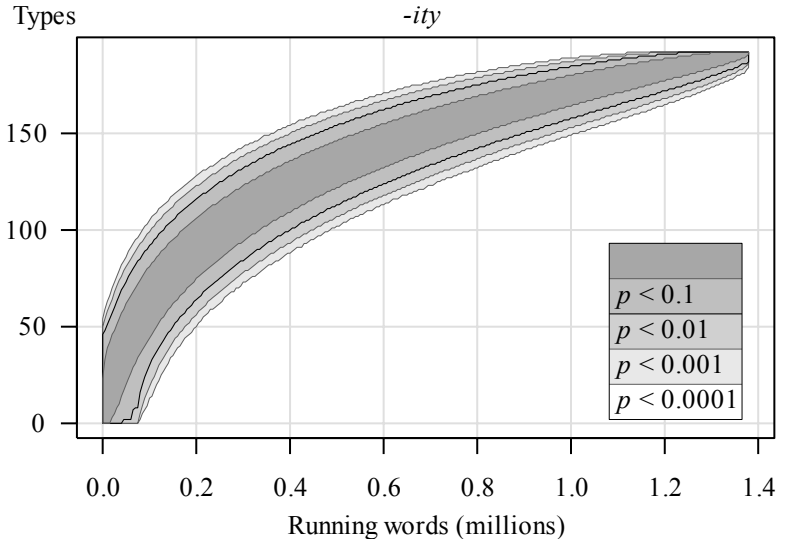

Figure 4. Bounds for -*ity* types as a function of the number of running words.

An example of this is shown in Figure 5. In the subcorpus which consists of the letters written by women, we have 340,116 running words and only 80 -*ity* types. In the subcorpus which consists of the letters written by men, we have 1,038,951 running words and 183 -*ity* types. We have plotted both data points on top of the curves already shown in Figure 4. We note that the data point which corresponds to women's texts lies below the lower bound with $p = 0.001$. We conclude that it is highly unlikely to come up with such a collection of samples by chance; our main hypothesis is true. We come back to the analysis of the results in Section 6.

As we shall see, calculating the curves requires some amount of computation. However, once we have done the computation, we can use the same curves repeatedly to answer various questions. We can test other similar hypotheses easily by plotting more data points on top of the curves. Indeed, we can do exploratory data analysis by plotting data points corresponding to each possible value of each sociolinguistic category, such as gender, domicile, social rank, and time period. We shall see examples of this in Section 6.

We can also analyse the curves qualitatively: the shape of Figure 4 above illustrates the nonlinear relation between the size of the subcorpus and the number of types occurring in it. Finally, we can calculate similar curves for measure (b), hapaxes, and we can also consider definition (ii), which means that the x axis shows the number of -*ity* tokens instead of the number of running words in the subcorpus. See Figure 6 for an example.

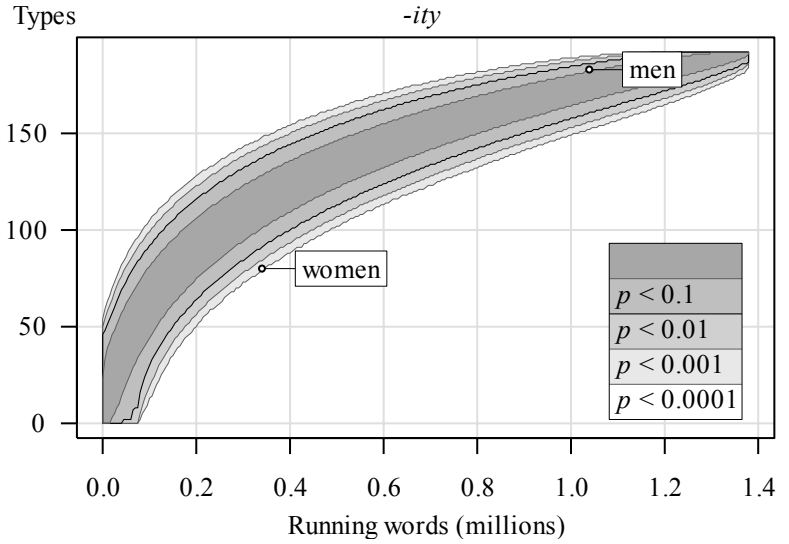

Figure 5. Hypothesis testing. Women have significantly few *-ity* types.

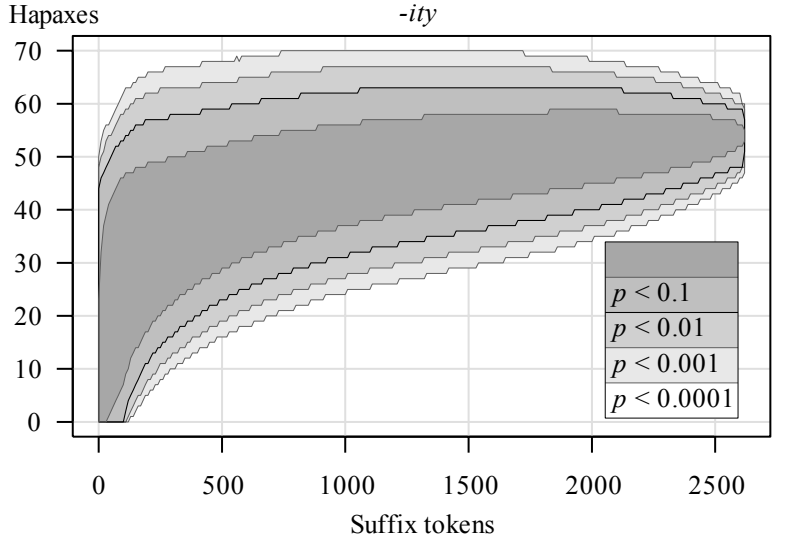

Figure 6. Bounds for *-ity* hapaxes as a function of the number of *-ity* tokens.

5.1 Basic algorithm

We proceed to present the operation of the computer program. The program performs the computations in two steps. The first step essentially tabulates for each pair (t, n) an approximation of the number of permutations such that there are exactly t types within the first n running words. The second step uses the table to find for each value of n those values of t at which we cross the significance levels of interest, such as $p = 0.01$ and $p = 0.001$.

The first step is computationally more intensive. It consists of generating a large number of random permutations of the samples – typically, the number of permutations is in the range of tens of thousands to millions. For each permutation, we process the samples one by one, in the order indicated by the permutation. For each new sample, we compute the total number of types observed so far. Each permutation can be interpreted as a type accumulation curve, similar to the two examples illustrated in Figure 7; in the figure, each tick mark corresponds to one sample. Once we have a complete accumulation curve, we increment the counters in the table for each pair (t, n) through which it passes. This is repeated for each permutation, after which we can perform the second step.

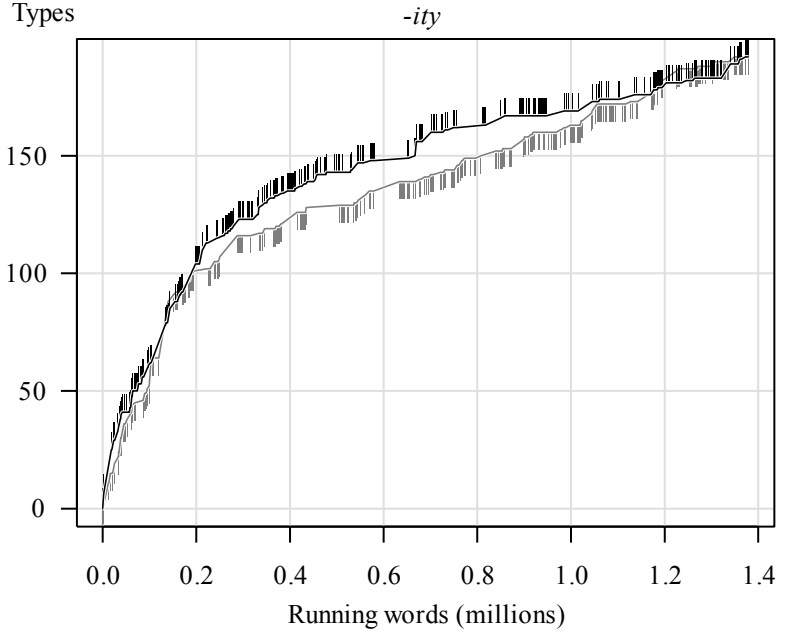

Figure 7. Two type accumulation curves. Each tick mark represents the addition of one sample.

5.2 Computational complexity

In the first step, we employ a randomised algorithm to approximate the number of permutations for each (t, n). This is an application of the Monte Carlo method (Mitzenmacher and Upfal 2005: 252), in which one picks a number of objects at random from a suitable probability distribution, checks which percentage of them satisfies the desired properties, and derives an estimate of the total number of such objects. By increasing the number of objects that we choose, we can improve the accuracy of the estimate. As is usual in an implementation of permutation testing (Good 2005: 233), we choose a particularly simple probability distribution, the uniform distribution over all permutations; therefore, we can pick a random permutation by using a simple algorithm for randomly shuffling a list.

By resorting to a randomised approximation algorithm, we have sacrificed some accuracy. This is acceptable, as we only need the first few decimals of the probability p. Approximation is in any case unavoidable, because it is not likely that there exists any efficient algorithm for, say, computing the exact number of permutations which traverse through a given point (t, n). Even determining whether the number is more than zero is hard: this is a generalisation of the SET COVER problem, which belongs to the class of NP-complete problems, and it is generally believed that no efficient algorithm exists for any problem that is NP-complete (see, e.g., Garey and Johnson 2003 [1979]).

5.3 Implementation details

Next we address the fact that we only have data at the granularity of entire samples. Put simply, based on our input data, we do not know whether the occurrences of the types are at the beginning or the end of the sample; if we are interested in knowing the exact value of t for some n which happens to be in the middle of a sample, we do not know whether we would have already met the new types of this sample by n running words or not. Therefore, our program adopts a safe approach: it always considers the worst case for us and the most favourable case to the null hypothesis, i.e., the case which produces the widest confidence intervals.

Finding the worst cases for the number of types is straightforward. For lower bounds, we can proceed as if all types were clustered at the very end of the sample, and for upper bounds we can assume the opposite. The case of hapaxes is more involved, as we need to distinguish between several cases: (a) newly created hapaxes, i.e., types which have not occurred before this sample and which occur only once in this sample; (b) temporary hapaxes, i.e., types which have not occurred before this sample and which occur more than once in this sample; and (c) removed hapaxes, i.e., types which have occurred exactly once before this sample and which occur at least once in this sample. For lower bounds, the worst case is that the types of class (c) occur at the very beginning of the sample, cancelling previously known hapaxes. For upper bounds, the worst case is that the

types of class (a) and one instance of each type of class (b) occur at the very beginning of the sample, increasing the number of hapaxes at least temporarily.

To develop a program which is computationally efficient in terms of time and memory requirements, we need to address some further issues. First, while the range of possible values of t is typically moderate, the range of possible values of n can be large; in our data, we have more than one million running words. The size of the table where the number of permutations for each (t, n) are stored would be impractical. We can significantly improve performance by dividing the n dimension into a smaller number of **slots**; for example, we can interpret the range from $n = 0$ to $n = 4{,}999$ as one slot, the following 5,000 running words as another slot, and so on. The approach of using slots is combined with the approach of finding worst-case bounds. Therefore, the slots can be used safely: they do not introduce any artefacts in the curves which would make some finding seem statistically significant if this is not the case. Naturally, using very large slots may prevent one from finding even statistically significant results.

To further improve performance, the computations in the first phase use a data layout in which each element requires only 1 or 2 bits of storage: for types, the single bit stands for "at least 1"; for hapaxes, one bit stands for "at least 1" and the other for "at least 2". The input is pre-processed into an incidence matrix which is stored in this compact format, and the table containing the counts for each slot is also stored in this manner. The compact memory layout is cache-friendly and allows us to exploit bit-parallelism in the calculations.

The program is written in standard C (ISO/IEC 9899:1999); it should compile and run on any standard-compliant platform. The only essential limitation on the size of the input data is the amount of available memory. Parameters such as the number of iterations and the slot size can be set by using command line switches.

5.4 Performance

The following example illustrates the typical performance of the program. In our input data for the suffix *-ity*, we had 412 samples and 192 different types of *-ity*. We used slots of 5,000 running words each; this resulted in 277 slots. We ran the experiments on a desktop PC with a 2.4-GHz Pentium 4 processor, under the Linux operating system; the application was compiled using the C compiler from the GNU Compiler Collection (GCC).

We experimented with two different numbers of permutations: 20,000, which is suitable for getting a quick idea of whether there are any statistically significant results in view, and 1,000,000, which is more than enough to produce publication-quality illustrations such as those presented in this work. The running time for computing the type accumulation curves was 1.7 seconds for 20,000 permutations and 82 seconds for 1,000,000 permutations. The running time for computing the

hapax accumulation curves was 2.3 seconds for 20,000 permutations and 113 seconds for 1,000,000 permutations.

5.5 Using the implementation

The computer program described in this section is freely available under an open source license (GNU General Public License, version 2.0 or later). For details on obtaining and using the program, see Suomela (2007).

Both the input and the output of the program are plain text files. The program accepts as input data matrices similar to those illustrated in Table 1. The input files can be prepared manually or, as we have done, by using corpus-specific tools. The output consists of the numerical data for curves similar to those in Figure 4 above. Tools such as statistical software packages or spreadsheets can be used to visualise the results. With our program, we provide a script which illustrates how to draw graphs similar to Figure 4 by using R, the free software environment for statistical computing (R Development Core Team 2007).

As stated above, the program is only needed for computing the upper and lower bounds for type accumulation, and such computation needs to be performed only once for a given data set. In the following section, we use the bounds for both hypothesis testing and exploratory data analysis.

6. Results and conclusions

Our hypothesis was that gender is significant in the case of *-ity*; as seen from Figure 5, this turned out to be the case. The richness of *-ity* types is significantly low ($p < 0.001$) in women's letters in the 17th-century part of the *CEEC*. Naturally, the 17th-century part of the *CEEC* is not a perfect representation of 17th-century English; neither are type counts a perfect measure of morphological productivity. Nevertheless, a result which is statistically this significant demands an explanation, and we argue that an attractive candidate can be found through examining the socio-historical situation in 17th-century England (see, e.g., Wrightson 1993). As women's access to education was severely restricted, they would not have had the competence to use the learned and etymologically foreign suffix *-ity* to the same extent as men.

The situation for *-ness* is shown in Figure 8. Here the data points for both men and women fall between the upper and lower bounds, and we cannot draw a similar conclusion on the significance of gender.

Finally, we explore some other sociolinguistic categories. Subcorpora based on the domiciles of the informants show no significant results. As for social rank, we might have expected to find a significantly low level of productivity for *-ity* in the lowest ranks, but there is simply too little data from them in the corpus.

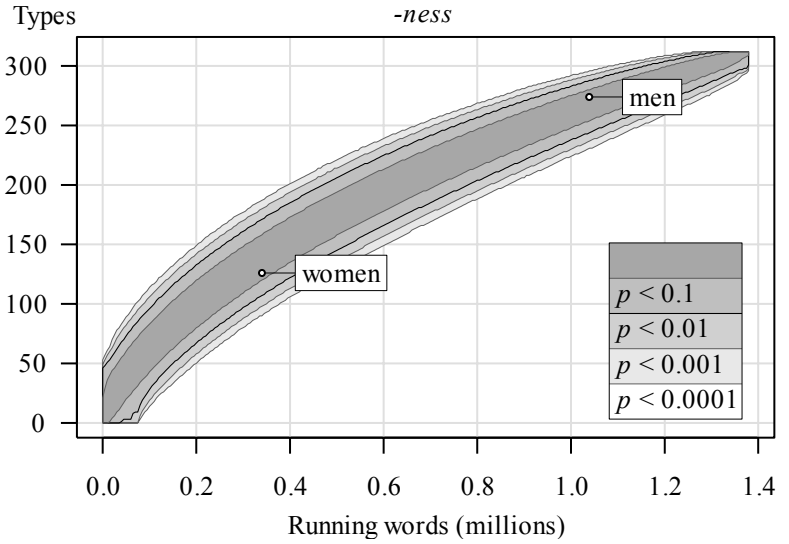

Figure 8. Bounds for *-ness* types as a function of the number of running words.

A more interesting case comes up when we divide the corpus into time periods: letters written in 1600–1639, and those written in 1640–1681. Figure 9, based on the same set of curves as Figure 5 above, shows that the type richness of *-ity* is significantly low in the earlier period. One interpretation for this could be that there is a linguistic change in progress: in the course of the 17th century, the use of *-ity* becomes more common in personal letters. This makes sense – not only was the use of Latinate features socially stratified (they were mostly used by learned men), but it was also register-specific, and began to spread from more formal contexts to less formal ones during the 16th and 17th centuries (cf. Nevalainen and Tieken-Boon van Ostade 2006: 281–282; Riddle 1985: 455–456).

The above examples illustrate the ease with which we can do exploratory data analysis once we have computed the bounds of the type accumulation curves. Even with a relatively small corpus, we were able to not only confirm our hypothesis but also discover unanticipated linguistically interesting results.

The bounds for hapax counts turned out to be too wide for significant differences to emerge (see Figure 6 above). It may be that this measure requires more data to become usable. However, if the problem of wide bounds for hapax accumulation curves persists in larger corpora, this could call into question the use of hapax-based productivity measures in general.

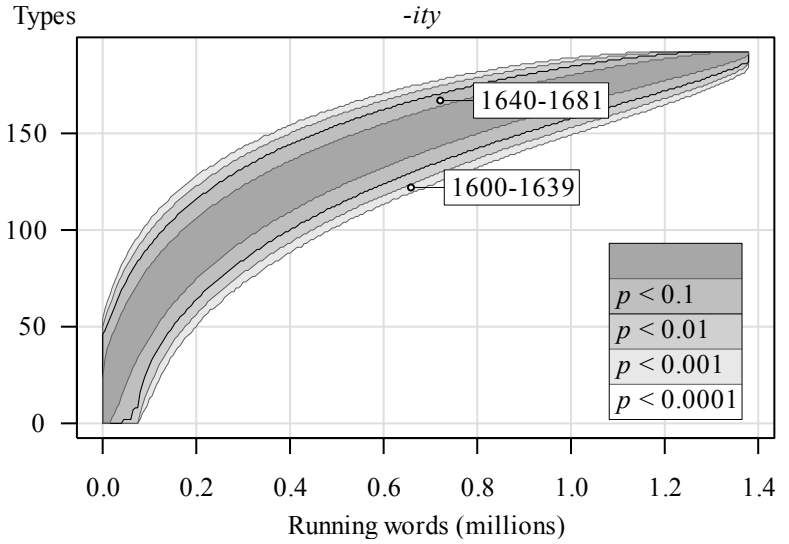

Figure 9. Subcorpora based on time periods.

In addition to testing hapax accumulation in larger corpora, future work could include a comparison between our type accumulation curves and those derived from more widely used parametric models. Another opportunity for future research would be a more fine-grained investigation of the differences between men and women in the use of the suffix *-ity*: as pointed out by an anonymous reviewer, part of the differences observed in this study could be due to women writing about a more restricted set of topics, which may lead to a large vocabulary overlap between women.

As noted in Section 4.1, our work focuses on definition (i) of corpus size – in our type accumulation curves, the *x* axis is the number of running words in the corpus. Another possibility would have been to compute type accumulation as a function of suffix tokens. Further work is needed in order to better understand the interplay between the number of running words, the number of affix tokens, and the number of affix types in the context of productivity.

Acknowledgements. We thank Harald Baayen, Terttu Nevalainen, the audience at ICAME 28 and the members of VARIENG for discussions and comments, and anonymous reviewers for their helpful feedback. The database of sociolinguistic information used in the study was compiled by Arja Nurmi. This research was supported in part by the Academy of Finland Centre of Excellence funding for the Research Unit for Variation, Contacts and Change in English (VARIENG) at the Department of English, University of Helsinki, and the Helsinki Graduate School in Computer Science and Engineering (Hecse).

Notes

1 As noted by an anonymous reviewer, this particular example could also be regarded as an instance of affix substitution. This provides an even stronger motivation for not leaving out these kinds of words.

References

Aronoff, M. and F. Anshen (1998), 'Morphology and the lexicon: Lexicalization and productivity', in: A. Spencer and A. M. Zwicky (eds.) *The Handbook of Morphology*. Cambridge, MA: Blackwell Publishers. 237–247.

Baayen, R. H. (1992), 'Quantitative aspects of morphological productivity', in: G. Booij and J. van Marle (eds.) *Yearbook of Morphology 1991*. Dordrecht: Kluwer Academic Publishers. 109–149.

Baayen, R. H. (1993), 'On frequency, transparency and productivity', in: G. Booij and J. van Marle (eds.) *Yearbook of Morphology 1992*. Dordrecht: Kluwer Academic Publishers. 181–208.

Baayen, R. H. (2001), *Word Frequency Distributions*. Dordrecht: Kluwer Academic Publishers.

Baayen, R. H. and R. Lieber (1991), 'Productivity and English derivation: A corpus-based study', *Linguistics*, 29: 801–843.

Baayen, R. H. and A. Renouf (1996), 'Chronicling the Times: Productive lexical innovations in an English newspaper', *Language*, 72 (1): 69–96.

Bolinger, D. L. (1948), 'On defining the morpheme', *Word*, 4: 18–23.

Březina, V. (2005), *The Development of the Prefixes* un- *and* in- *in Early Modern English with Special Regard to the Sociolinguistic Background*, unpublished MA thesis, Faculty of Arts, Charles University in Prague.

CEEC = Corpus of Early English Correspondence (1998), compiled by the Sociolinguistics and Language History project team (T. Nevalainen, J. Keränen, M. Nevala, A. Nurmi, M. Palander-Collin, H. Raumolin-Brunberg) at the Department of English, University of Helsinki. <http://www.helsinki.fi/varieng/domains/CEEC.html>.

Corpus of Early English Correspondence Sampler (1998), see above.

Cowie, C. and C. Dalton-Puffer (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*. Amsterdam: Rodopi. 410–437.

Dalton-Puffer, C. (1996), *The French Influence on Middle English Morphology: A Corpus-Based Study of Derivation*. Berlin: Mouton de Gruyter.

Evert, S. and M. Baroni (2005), 'Testing the extrapolation quality of word frequency models', in: P. Danielsson and M. Wagenmakers (eds.), *Proceedings of Corpus Linguistics 2005*. The Corpus Linguistics Conference Series 1. Available at <http://www.corpus.bham.ac.uk/PCLC/>.

Evert, S. and M. Baroni (2007), '*zipfR*: Word frequency distributions in R', in: *Proceedings of the ACL 2007 Demo and Poster Sessions*. Stroudsburg, PA: Association for Computational Linguistics. 29–32.

Garey, M. R. and D. S. Johnson (2003) [1979], *Computers and Intractability: A Guide to the Theory of NP-Completeness*. New York: W. H. Freeman and Company.

Good, P. (2005), *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. 3rd edition. Springer Series in Statistics. Berlin: Springer-Verlag.

Gotelli, J. and R. Colwell (2001), 'Quantifying biodiversity: Procedures and pitfalls in the measurement and comparison of species richness', *Ecology Letters*, 4: 379–391.

Hay, J. (2001), 'Lexical frequency in morphology: Is everything relative?', *Linguistics*, 39 (6): 1041–1070.

Marchand, H. (1969), *The Categories and Types of Present-Day English Word-Formation: A Synchronic-Diachronic Approach*. 2nd edition. Munich: C. H. Beck'sche Verlagsbuchhandlung.

MED = Middle English Dictionary, 2001 edition. Electronic version. Available at <http://ets.umdl.umich.edu/m/med/>.

Mitzenmacher, M. and E. Upfal (2005), *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge: Cambridge University Press.

Nevalainen, T. and H. Raumolin-Brunberg (2003), *Historical Sociolinguistics: Language Change in Tudor and Stuart England*. London: Pearson Education.

Nevalainen, T. and I. Tieken-Boon van Ostade (2006), 'Standardisation', in: R. M. Hogg and D. Denison (eds.) *A History of the English Language*. Cambridge: Cambridge University Press. 271–311.

OED = Oxford English Dictionary, 2nd edition, 1989. OED Online. Available at <http://dictionary.oed.com>.

Plag, I. (1999), *Morphological Productivity: Structural Constraints in English Derivation*. Berlin: Mouton de Gruyter.

R Development Core Team (2007), *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at <http://www.R-project.org>.

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*. Berlin: Mouton de Gruyter. 435–461.

Säily, T. (2005), 'Use of the suffixes *-ity* and *-ness* in early English letters: Was gender a factor?', unpublished seminar paper, Department of English, University of Helsinki.

Säily, T. (2008), *Productivity of the Suffixes* -ness *and* -ity *in 17th-century English Letters: A Sociolinguistic Approach*, unpublished MA thesis, Department of English, University of Helsinki.

Säily, T. and J. Suomela (2007), 'Incidence matrices for *-ness* and *-ity*'. Available at <http://www.cs.helsinki.fi/jukka.suomela/ity-ness-data/>.

Suomela, J. (2007), 'Type and hapax accumulation curves', computer program. Available at <http://www.cs.helsinki.fi/jukka.suomela/types/>.

Tweedie, F. J. and R. H. Baayen (1998), 'How variable may a constant be? Measures of lexical richness in perspective', *Computers and the Humanities*, 32: 323–352.

Wrightson, K. (1993), *English Society, 1580–1680*. London: Routledge.