Variation in morphological productivity in the BNC: Sociolinguistic and methodological considerations

TANJA SÄILY*

Abstract

The first aim of this work is to examine gender-based variation in the productivity of the nominal suffixes -ness *and* -ity *in present-day British English. Possible interpretations are presented for the findings that* -ity *is used less productively by women, while with* -ness *there is no gender difference. The second aim is to analyse the validity of hapax-based measures of productivity in sociolinguistic research. It is discovered that they require a significantly larger corpus than type-based ones, and that the category-conditioned degree of productivity* P *is unusable when comparing subcorpora based on social groups. Otherwise, hapax legomena remain a theoretically well-founded component of productivity measures.*

Keywords: *sociolinguistics, gender variation, present-day English, methodology, morphological productivity, word-formation, nominal suffixes, hapax legomena*

1. Introduction

According to Adamson (1989: 204), English is remarkable for its "double lexicon", in which almost all native words have Romance or Latinate synonyms. She argues that this can be seen as a case of diglossia, a sociolinguistic situation in which there is a dual standard of literary and colloquial norms (Adamson 1989: 205–207). The phenomenon is not restricted to the lexicon, however, many derivational affixes of Germanic origin having a synonymous double of French or Latin origin. One of these pairs is formed by the nominal suffixes *-ness* and *-ity*, which have featured prominently in studies of English word-formation (e.g., Aronoff 1976; Anshen and Aronoff 1989; Plag et al. 1999; Hay and Baayen 2003). Nevertheless, social aspects of the use of these suffixes have not been addressed satisfactorily, which is surprising considering their

Corpus Linguistics and Linguistic Theory 7–1 (2011), 119–141 1613-7027/11/0007–0119
DOI 10.1515/CLLT.2011.006

diglossic nature. In fact, the question of sociolinguistic variation in morphological productivity in general has received very little attention thus far, the study of Dutch affixes by Keune et al. (2006) being a rare exception.

The present work is a quantitative, corpus-linguistic study of sociolinguistic variation in the morphological productivity of the suffixes *-ness* and *-ity*, which are typically used to derive abstract nouns from adjectives, as in (1).

(1) *prescriptive* + *-ness* → *prescriptiveness*
prescriptive + *-ity* → *prescriptivity*

As mentioned above, the suffixes are roughly synonymous, meaning something like 'the state or quality of being ADJ' (although some semantic differences have been noted by scholars such as Riddle 1985). While *-ness* is a native suffix, *-ity* was borrowed from French during the Middle English period and was later reinforced through Latin in Early Modern English. It is precisely this difference in their origins that makes the pair interesting from the point of view of sociolinguistics. Romaine (1985: 461–462) argues that the borrowed suffixes were initially only available to highly educated individuals, who were most often men. This inequality seems to have led to stylistic and situational differences in the use of *-ity* and *-ness* that persist to this day. In other words, *-ness* and *-ity* became part of the diglossic situation in English as described by Adamson (1989), *-ity* being the 'high' variety synonym and *-ness* the 'low' variety one.

Even in present-day English, *-ity* is more selective in that it is almost exclusively attached to bases of a French or Latin etymology, whereas *-ness* can be freely attached to both native and foreign bases (Marchand 1969: 312, 334). Because of this and the above-mentioned semantic differences, the suffixes are not entirely interchangeable, which creates a problem if they are to be studied within the variationist sociolinguistics framework. That is, they do not constitute a perfect linguistic variable, defined by Milroy and Gordon (2003: 88) as a linguistic item with variant realisations which refer to the same thing but which covary with different items or social categories. This problem can be overcome by analysing the variation in the productivity of each suffix separately, looking at their frequency of use by various social groups (see, e.g., Nevalainen 2006: 357).

Following up on a study of these suffixes in 17th-century data in which a clear gender difference emerged (Säily and Suomela 2009), the sociolinguistic question examined here is whether men and women use these suffixes differently in present-day English, i.e., whether there is gender-based variation in the morphological productivity of the suffixes. Furthermore, the study seeks to answer a related question pertaining to corpus-linguistic methodology. Some of the most commonly used measures of morphological productivity (e.g., Baayen 1993) are based on words occurring only once in the corpus, or hapax

legomena. In the 17th-century study, however, hapaxes proved unusable, perhaps owing to the small size of the corpus. The present work, which uses a much larger corpus, aims to find out whether hapax-based productivity measures can be considered valid in sociolinguistic research.

The motivation for the sociolinguistic question is explained in the next section. The question of the validity of hapax-based productivity measures, along with its theoretical background, is explored in more detail in Section 3. Section 4 introduces the material used in this study, the *British National Corpus* (BNC). Section 5 presents the sociolinguistic and methodological results, which are discussed further in Section 6. Finally, Section 7 presents concluding remarks.

2. Background: Gender and variation

Without a doubt, gender is one of the most important categories in present-day sociolinguistics. The term gender is used instead of sex to emphasise the social nature of the concept (e.g., Nevalainen and Raumolin-Brunberg 2003: 110). Gender roles can change with societal norms, which in turn can affect the ways in which men and women use language in their daily lives. Nevertheless, study after study has shown that it is women who tend to be the leaders of language change – as Tagliamonte and D'Arcy (2009: 63) point out, "one of the most consistent findings of sociolinguistic research has been the gender-asymmetric nature of the process". Consistent gender differences have emerged in synchronic variation as well, to the point that this has been called a sociolinguistic fact (Hudson 1996: 202; Nevalainen 2006: 360–361).

As noted by scholars such as Cameron (2006: 734), there has been an ongoing shift in gender research from difference to diversity since the 1990s. The study of binary differences between men and women has given way to an awareness of internal variation among men on the one hand and women on the other, with an emphasis on how masculinity and femininity are constructed in different contexts. Some scholars in fact reject the search for gender differences altogether as an "impoverished framework" (Talbot 2006: 741). Many, however, see a need for the coexistence of multiple approaches in order to get a more complete picture (e.g., Holmes and Meyerhoff 2003). Different frameworks focus on different questions, and the study of the local needs to be complemented by the study of the global (Cameron 2006: 738–739). For instance, identifying gender differences may become necessary if we wish to compare the influence of gender on the English language in the present and the past (e.g., Cameron 2008).

The role of corpus linguistics, of course, is crucial in the study of the past, because we do not have direct access to the informants. Even for studies of present-day language, however, corpus linguistics can offer a variety of

quantitative methods and large amounts of naturally-occurring language data. Quantitative work makes for generalisable results – something that qualitative gender research alone cannot produce, although naturally both are necessary. Furthermore, the time depth of many changes is longer than can be studied through apparent-time methods, which argues for the use of diachronic corpora and comparisons between present-day and historical materials (e.g., Nevalainen and Raumolin-Brunberg 2003; Nurmi et al. 2009). As an example, the extremely intriguing results by Tagliamonte and D'Arcy (2009; following up on Labov 2001: Ch. 14) that men lag a generation behind women in language change would certainly benefit from real-time confirmation.

Gender continues to be used as one robust social category among others in variationist sociolinguistics (see, e.g., Labov 2001: 262), but the focus nowadays is more on multivariate analyses that take into account both interaction across categories and variation within categories (e.g., Tagliamonte and D'Arcy 2007). Furthermore, it is still universally acknowledged that sociolinguistic variation is probabilistic rather than categorical (Cameron 2006: 733–734; Nevalainen 2006: 358–359). While, for instance, there is a broad trend for women to use more personal pronouns than men (Rayson et al. 1997), it is of course possible to find many individuals who differ from this norm for various reasons, such as their social background or their aims in the particular discourse situation. Nevertheless, these broad trends can be interpreted as part of gendered discourse styles, for which an increasing amount of evidence is becoming available (e.g., Holmes 1998; Palander-Collin 1999; Biber and Burges 2000).

The present work can be placed within the framework of quantitative sociolinguistics. While the study concentrates on an observed difference between men's and women's language use, it is by no means blind to diversity within the social categories of 'man' and 'woman'. The starting-point of my research was exploratory data analysis using all of the sociolinguistic metadata provided with the corpus; combinations of categories, such as gender and social class, were also created and tested. As in so many previous studies (cf. Hudson 1996: 202; Nevalainen 2006: 360–361), the gender difference emerged from the data.

The impetus for this study was provided by previous work on the same suffixes in the 17th-century part of the *Corpus of Early English Correspondence* (CEEC; Säily and Suomela 2009), research which showed that the productivity of *-ity* was significantly lower in letters written by women, while there was no significant variation in the use of *-ness*. This was explained by the fact that *-ity* was a 'learned' and etymologically foreign suffix, which at the time could have been used competently only by those with a classical education, i.e., by high-ranking men (unfortunately, there was too little data from the lower ranks to confirm the influence of socio-economic status).

A related study, however, found that women consistently used fewer nouns than men in the corpus, while the opposite held for personal pronouns (Säily et al. forthcoming; cf. Rayson et al. 1997 and Argamon et al. 2003 for similar results in present-day English). Thus, the result that women used the nominal suffix *-ity* less diversely than men might in fact be expected, because women used fewer nouns overall. The question then becomes why women did not use *-ness* significantly less than men.[1] A comparison with present-day English might be useful here, providing an additional motive for the present work.

3. Measuring productivity in corpus linguistics

Productivity has long been seen as one of the most problematic issues in derivational morphology (see, e.g., Kastovsky 1986). A quantitatively oriented definition of the phenomenon was offered as early as Bolinger (1948: 18): "the statistically determinable readiness with which an element enters into new combinations". Exactly how this readiness should be determined, however, has been the subject of much debate.

An easily accessible measure of productivity is type frequency, i.e., the number of different words of a particular morphological category found in a corpus. According to Dalton-Puffer (1996: 217), type frequency has an obvious connection with morphological productivity: "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". Nevertheless, there are some problems with this measure – for instance, a large number of types could indicate past rather than current productivity, as the words in question could have been in the language for centuries (Cowie and Dalton-Puffer 2002: 416).

A concept related to type frequency is token frequency, which is the number of all words of a particular morphological category found in a corpus. Each time a word belonging to the category is encountered, it is counted, regardless of whether the same word has occurred before. Token frequency alone cannot be used as an indicator of productivity, because "token count is often inflated by a small number of very common types" (Cowie and Dalton-Puffer 2002: 426).

In the early 1990s, Harald Baayen and colleagues began to develop a set of corpus-linguistic measures intended to capture different facets of the elusive concept of morphological productivity (e.g., Baayen and Lieber 1991; Baayen 1992; Baayen 1993). Some of these are still being used and are recommended by, e.g., Baayen (2008). They are based on the frequencies of types, tokens and hapax legomena (or hapaxes), the third term referring to words occurring only once in the corpus. Hapax frequency, then, means the number of words of a particular morphological category occurring only once in a corpus.

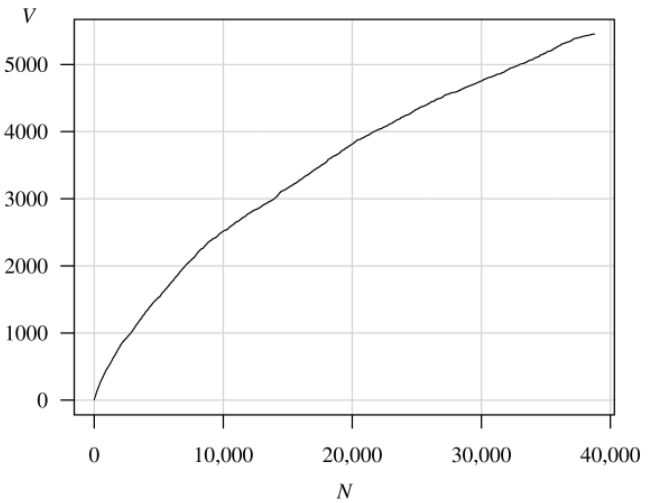

Figure 1. *The type accumulation curve for all types* V *as a function of all tokens* N *in the Project Gutenberg e-text of Joseph Conrad's* Heart of Darkness*, <http://www.gutenberg.org/2/1/219/>. Reproduced from Säily (2008: 15)*

The first of Baayen's measures is the extent of use V, which simply refers to type frequency as defined above. It can be shown that type frequency grows as a function of token frequency, but in a non-linear fashion (Baayen 1992: 113), as in Figure 1. This means that the type frequencies of different affixes, or of the same affix in different subcorpora, cannot be compared directly unless their token frequencies are of a similar magnitude. Normalising type frequencies is not an option because normalisation presupposes linearity: a simple division of the number of types by the number of tokens is equivalent to drawing a line from the origin to the endpoint of the curve in Figure 1. This is clearly not an adequate representation of the curve and will lead to trouble if comparisons are made at any other point of token frequency than the endpoint.

Baayen's second measure is the category-conditioned degree of productivity P, which is defined as hapax frequency divided by token frequency: $P = n_1/N$. Baayen (1992: 115–117) shows that this approximates the growth rate of the vocabulary of the morphological category in question; i.e., the probability of encountering new types. In Figure 1, P could be drawn as the tangent to the endpoint of the curve.

The third measure is the hapax-conditioned degree of productivity P^*, which is defined as hapax frequency divided by the number of all hapaxes in a corpus: $P^* = n_1/h$. Baayen (1993: 192–193) argues that this estimates "the relative contribution of a given morphological category to the overall vocabulary growth", i.e., the probability that a new type represents a given morpho-

logical category. According to Hay and Baayen (2003: 101), when the comparison is between affixes within the same corpus, this can be simplified into hapax frequency alone: $P^* = n_1$.

All three measures are dependent on the size of the corpus, both in terms of token frequency and the number of running words in the corpus (see, e.g., Baayen and Lieber 1991: 817, 820; Baayen 1992: 113; Baayen 1993: 191). This makes comparisons between, e.g., *P* figures problematic. Gaeta and Ricca (2006) propose an improvement on *P* such that hapax frequencies are sampled at the same token frequency for each affix, but this entails taking samples from the corpus based on the token frequency of the least frequent affix, which discards valuable data and requires a large uniform corpus.

Säily and Suomela (2009) suggest another solution, based on accumulation curves and permutation testing. In this method, the corpus is divided into samples that are sufficiently large to preserve discourse structure (in the written component of the BNC, individual texts; in the spoken component, an individual person's contributions to the conversation). These samples are then taken in a random order to construct accumulation curves for a morphological category in the corpus. This can be done for both types and hapaxes, plotting them as a function of either token frequency or the number of running words in the corpus. The procedure is as follows.

Pick a sample randomly, calculate the number of types or hapaxes it contains, and plot it on a figure similar to Figure 1, with the number of types or hapaxes on the *y* axis and the number of suffix tokens or running words on the *x* axis. Pick another sample, add it to the previous one, calculate the number of types or hapaxes, and plot it on the figure. Repeat until the entire corpus has been sampled. This will produce something like Figure 2, which is essentially the same thing as Figure 1, with the exception that the corpus was processed sample by sample in random order rather than word by word in the original order.

The procedure of building accumulation curves for random permutations of the corpus is then repeated, say, a million times. After this, it is a relatively simple task to draw significance levels on the plot to indicate the area covered by, say, 99.9% of the curves. All this can be done automatically by a computer program (Suomela 2007). The next step is to plot the type or hapax frequencies of the desired subcorpora on the curves, which immediately shows whether a subcorpus is statistically significantly different in type or hapax frequency from the corpus as a whole. The result is an assumption-free, highly visual way of comparing the productivity of a morphological category in a subcorpus with the entire corpus, as in Figure 3.

The results shown in Figure 3 were discussed in Section 2. As can be seen from the figure, these results were obtained with type frequency as a function of the number of running words in the corpus. It would also have been possible to use type frequency as a function of the token frequency of the suffix. The

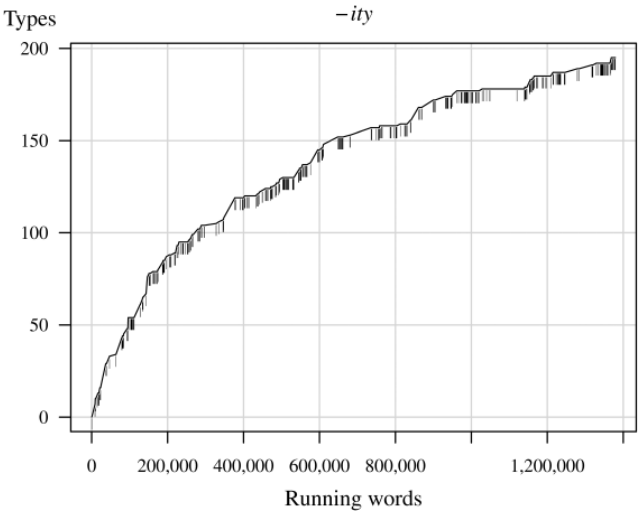

Figure 2. *A randomly constructed type accumulation curve for the suffix* -ity *in the 17th-century part of the CEEC. Each tick mark represents the addition of one sample. Based on Säily (2008: 68)*

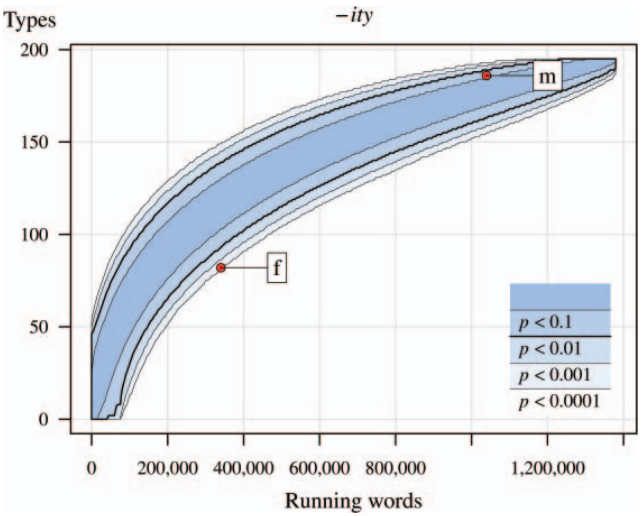

Figure 3. *Bounds for 1,000,000 type accumulation curves, with gender-based subcorpora (m = male, f = female) plotted on the curves, for the suffix* -ity *in the 17th-century part of the CEEC. Women have a significantly low type frequency. Based on Säily and Suomela (2009: Figure 5)*

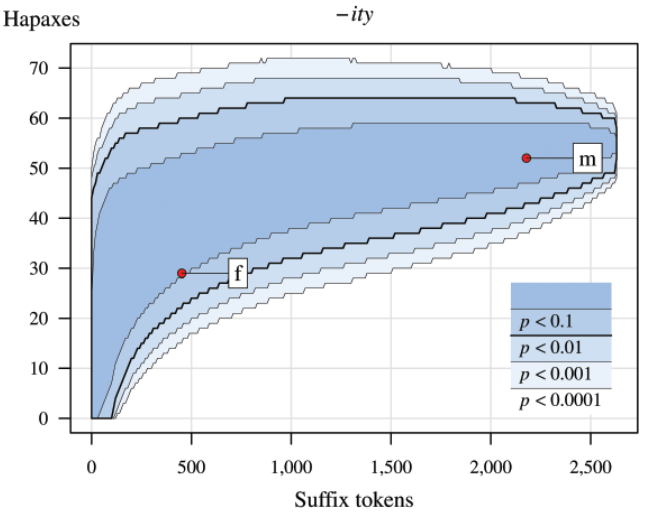

Figure 4. *Bounds for 1,000,000 hapax accumulation curves, with gender-based subcorpora plotted on the curves, for the suffix* -ity *in the 17th-century part of the CEEC. Based on Säily and Suomela (2009: Figure 6)*

number of running words was selected because this bore the most resemblance to the use of type frequency as a measure of morphological productivity in previous research, i.e., comparing type frequencies within the same corpus or across subcorpora of the same size in running words (as in Dalton-Puffer 1996: 106).

Hapax accumulation curves, however, proved unusable in the 17th-century study. Figure 4 illustrates why: the confidence intervals became so wide that no significant results were obtained. In fact, it seemed to be almost a matter of chance what number of hapaxes occurred in a corpus of a given size. This is worrying considering that hapax legomena are instrumental in two common measures of morphological productivity, *P* and *P**. Since the problem may arise from too little data, the present study will test hapax accumulation curves in a much larger data set, namely, the written part of the BNC.

With hapax frequency n_1 on the *y* axis and token frequency *N* on the *x* axis, Figure 4 can be regarded as a sort of equivalent of $P = n_1/N$. However, while *P* assumes that hapax frequency grows linearly with token frequency, Figure 4 shows that this is not the case – hapax accumulation curves are, after all, curves. This is another issue to be discussed in this paper.

4. Material

This study uses the *British National Corpus* (BNC), a 100-million-word corpus of British English compiled in the early 1990s. About 90% of the material

consists of written English of various genres. The rest is spoken, divided into two subcorpora: a demographically sampled component, chiefly comprising conversations, and a context-governed component containing spoken language from various situations, some quite formal.

Of the BNC subcorpora, the 4.2-million-word demographically sampled spoken component is the best suited for sociolinguistic research. This was compiled by randomly sampling people according to their region, age, gender and socio-economic status (Burnard 2007: 1.5.1). The respondents were given portable tape recorders and were asked to record all their conversations for a period of 2–7 days. They were also asked to fill out a form about each conversation, giving information on the time, participants and the nature of the situation (Burnard 2007: 1.5.1.1). The conversations were then transcribed orthographically, and the metadata was digitised. Since the corpus does not include audio files, researchers have to rely on the transcriptions.

Because of how the information was collected, a great deal is known about the respondents themselves, but the other speakers are less well documented, information on their dialectal background being especially patchy. Nevertheless, the gender of the speaker is known for about 88% of the data, and both social class and gender for 62% (2.6 million words). This 2.6-million-word subcorpus (henceforth called BNC-DS) is what was used as the spoken material for the present study.

The written component of the BNC contains 88 million words of written English from the 1960s to the early 1990s, the bulk of the samples coming from the period 1985–1993 (Burnard 2007: 1.3). Although the gender of the writer was not one of the sampling criteria, which mainly considered genre-related issues (Burnard 2007: 1.4.2), it is known for c. 51% of the data, or 45 million words. This is the subcorpus (henceforth called BNC-W) used as the written material for the present study. It can be further subdivided according to the "domain" of the texts (Burnard 2007: 1.4.2.3); this study utilises a rough division between imaginative (BNC-W_{imag}) and informative (BNC-W_{inf}) texts.

Although the BNC has been lemmatised and part-of-speech tagged, it became clear early on that the annotation could not be relied on to find all instances of the suffixes, and search queries were instead made based on the ending alone. The *BNCweb* software (see Hoffmann et al. 2008) was used to make the queries: *ness|*nesses for *-ness*, and *ity|*ities|*ety|*eties for *-ity*. The search hits were then pruned to include only genuine instances of the suffixes. The criteria were in the first instance etymological; for a justification of this, see Säily and Suomela (2009: 90). For instance, *business* was accepted as a *-ness* word, but *governess* was not; *sanctity* qualified as an *-ity* word, but *slappity* (as in *slappity slap*) did not. In the spoken data, further restrictions were imposed on what was counted in an equivalent manner to Säily (2008: 87–95). As in the 2008 study, since these restrictions proved to have no effect

on the main findings, the results presented here are based on data pruned using etymological criteria alone.

5. Results

5.1 *Spoken data, BNC-DS*

In BNC-DS, the productivity of both *-ity* and *-ness* turned out to be significantly lower in women's speech, as can be seen in Figures 5 and 6 respectively. As noted in Section 2, this was perhaps to be expected, because women have been shown to use fewer nouns overall. Of course, the frequency of nouns is measured in token frequency, while productivity pertains to type frequency, so they are not directly comparable. It turned out, however, that women used *-ness* and *-ity* less in terms of token frequency as well as type frequency; i.e., they used the suffixes both less often and less diversely than men.

For *-ity*, the difference is only just significant (Figure 5). Figure 6 shows that with *-ness*, gender difference is tied to social class: the difference is significant for lower-class women (from casual workers to skilled manual workers, BNC categories DE and C2), rather than women in general. Men of lower socio-economic status use *-ness* quite diversely, as in (2).

(2) *I always think there's a little <pause> shade of* ***big headedness*** *with Yul Bryner!* (BNC-DS: KBB 642)[2]

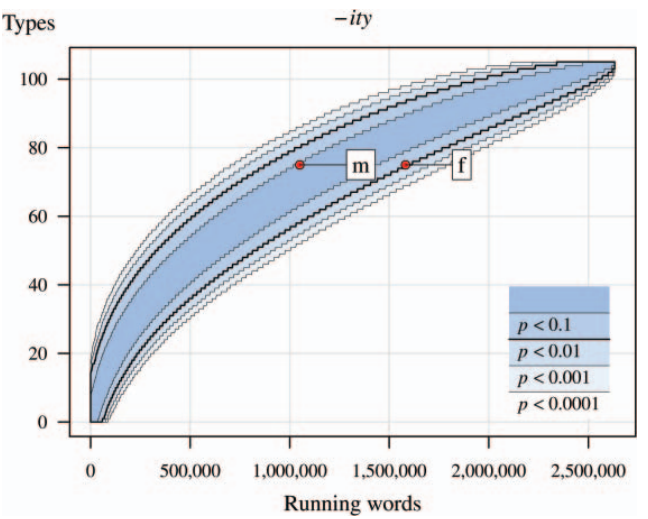

Figure 5. *Gender and* -ity *types in BNC-DS*

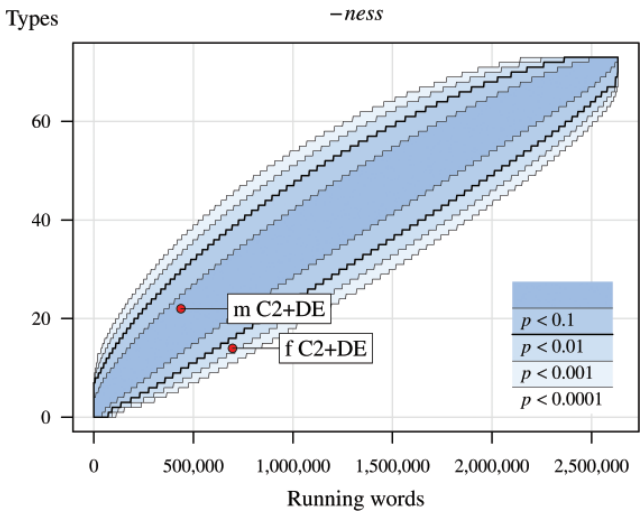

Figure 6. *Gender and* -ness *types in BNC-DS*

Women of the same status, on the other hand, almost always use lexicalised (or pragmaticised) *-ness* words in conversations, as in (3).

(3) *I just said,* ***goodness****, I said we've been waiting for this parcel since before Christmas and he says oh have you duck.* (BNC-DS: KDW 5280)

5.2 *Written data, BNC-W*

In BNC-W, the productivity of *-ity*, but not of *-ness*, proved significantly low in women's writing. Because this subcorpus contained more genres than the spoken subcorpus, it was conceivable that the result could have been distorted by genre variation – the bulk of men's writing belonged to the informative domain, while most of women's writing was to be found under the imaginative domain. The analysis was thus repeated for each domain separately; however, the result remained the same, as can be seen in Figures 7–10.

This result is interesting in that it mirrors the 17th-century one (Säily & Suomela 2009) discussed in Section 2. However, the tentative explanation offered for the 17th century does not apply here since women are no longer excluded from higher education, and *-ity* is no longer such a novel suffix that its use would necessarily require a classical education. The explanation relying on the overall low frequency of nouns, on the other hand, fails to address the question of why only *-ity* and not *-ness* is used significantly less productively. While the lack of a statistically significant difference does not conclusively prove that there is no difference, it seems that with such a large amount of data, if there

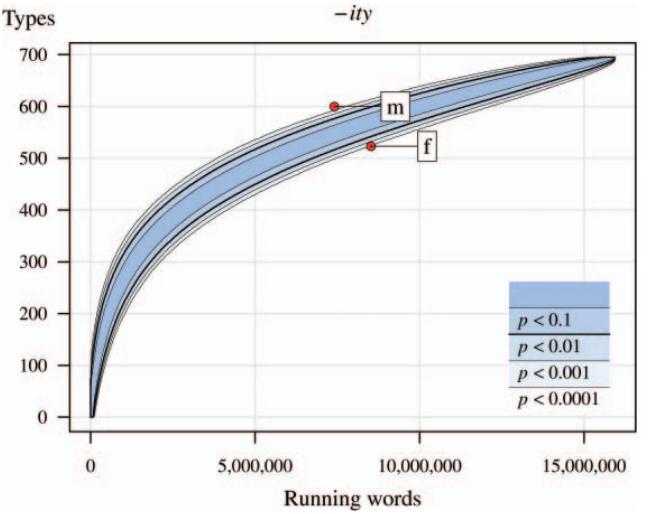

Figure 7. *Gender and* -ity *types in BNC-*W_{imag}

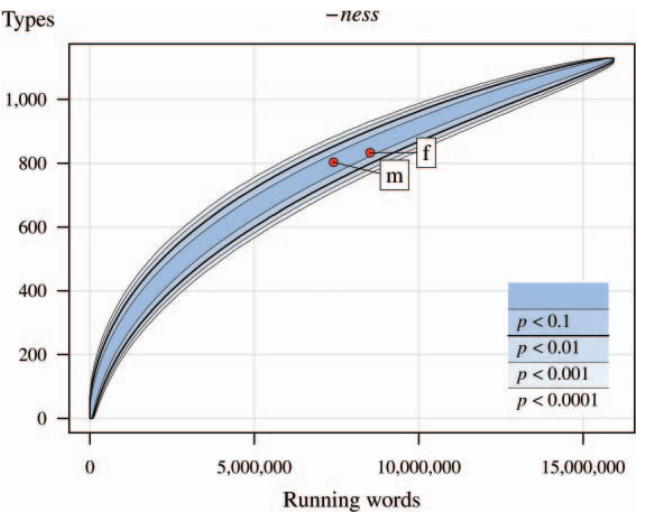

Figure 8. *Gender and* -ness *types in BNC-*W_{imag}

was a genuine difference, it would have emerged as statistically significant. Why, then, do women find *-ness* such a useful nominal suffix, when they are not very noun-oriented in general?

As noted in Section 1, *-ness* and *-ity* are not perfectly synonymous suffixes. Riddle (1985: 437) argues that "*-ness* tends to denote an embodied attribute or

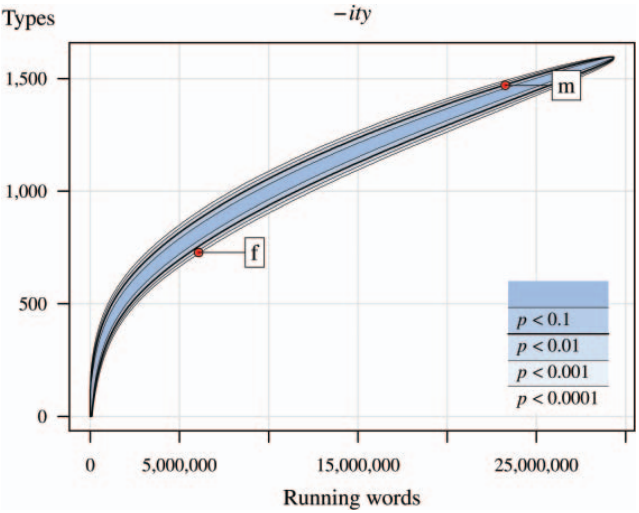

Figure 9. *Gender and* -ity *types in BNC-W_{inf}*

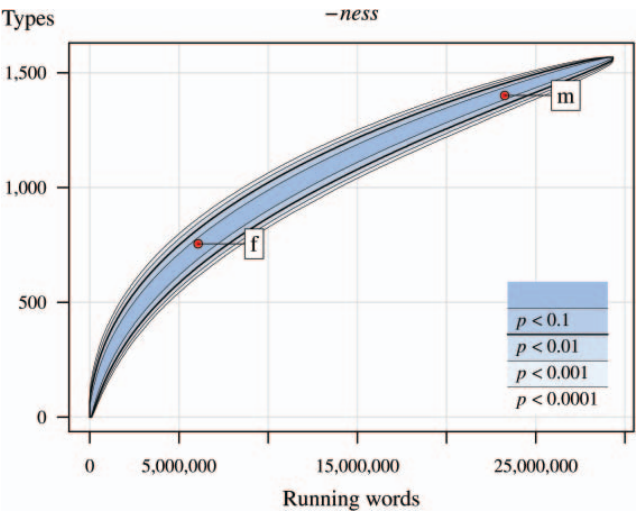

Figure 10. *Gender and* -ness *types in BNC-W_{inf}*

trait, while *-ity* tends to denote an abstract or concrete entity". Perhaps the meaning 'embodied attribute or trait' is well suited to women's writing style, which has been described as involved rather than informational (e.g., Argamon et al. 2003). This distinction goes back to Biber's multidimensional analysis of

register variation (e.g., 1988), in which he considers the co-occurrence patterns of a number of linguistic features in texts. One of the dimensions, identified through factor analysis, is labelled "Informational versus Involved Production". Among other things, the involved side of the dimension is characterised by a high frequency of personal pronouns, especially first- and second-person pronouns. Other features include private verbs, present-tense verbs and general emphatics. Many of these "mark interpersonal interaction or expression of personal feelings" (Biber and Finegan 1989: 491). Thus, an involved style might well prompt the use of *-ness* to describe personal traits.

As mentioned above, personal pronouns are a frequent feature of the involved style, and *-ness* seems to be often used with pronouns, as in example (4) by a female author. A search of the written component of the BNC (first for *ness_{N} and *ity_{N}, then for the same preceded by *_DPS) offers some support to this intuition: *-ness* tokens are used with possessive personal pronouns in c. 7.7% of the instances (women 11.6%, men 7.6%), while for *-ity* the figure is only 3.8% (women 6.2%, men 4.1%; the lower overall percentage is due to texts for which the author's gender is unknown). A similar tendency is observable in the demographically sampled spoken component of the corpus. Further research is needed to verify these initial findings and to determine whether this applies to types as well as tokens.

(4) *Most of them are common sense but you would be surprised how, in **our eagerness** to succeed, we often forget them.* (BNC-W_{inf}: AYK 245)

The entity meaning of *-ity* is illustrated in example (5) by a male author. Clearly, this has little to do with personal/interpersonal aspects of language use.

(5) *Other structures include **exclusivity**, where participation in one relationship excludes participation in another or **inclusivity**, where participation in one relationship automatically includes participation in another.* (BNC-W_{inf}: HRK 810 [emphasis original])

These results will be discussed further in Section 6.

5.3 *Productivity measures*

As described in Section 3, BNC-W was also used as a testbed for hapax-based productivity measures. The results in Sections 5.1 and 5.2 were obtained using type accumulation curves; equivalent results using hapax accumulation curves for *-ity* in BNC-W_{inf} are shown in Figures 11 and 12 (these correspond to Figure 9 in Section 5.2).

It is immediately obvious from the figures that the confidence intervals for hapax accumulation curves have indeed become narrower with a greater

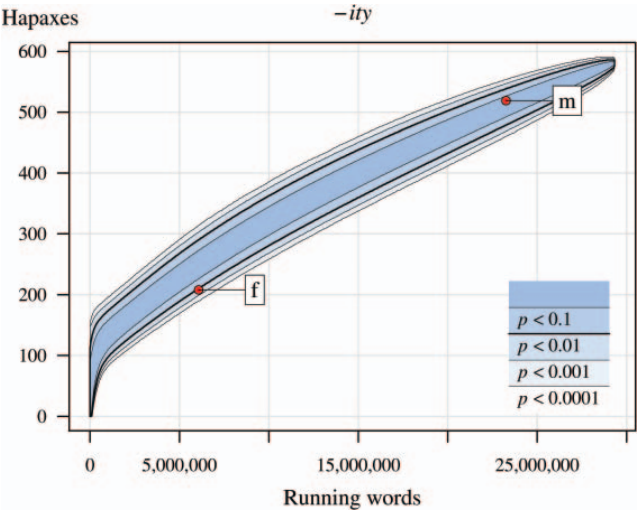

Figure 11. *Gender and* -ity *hapaxes as a function of the number of running words in BNC-W_{inf}*

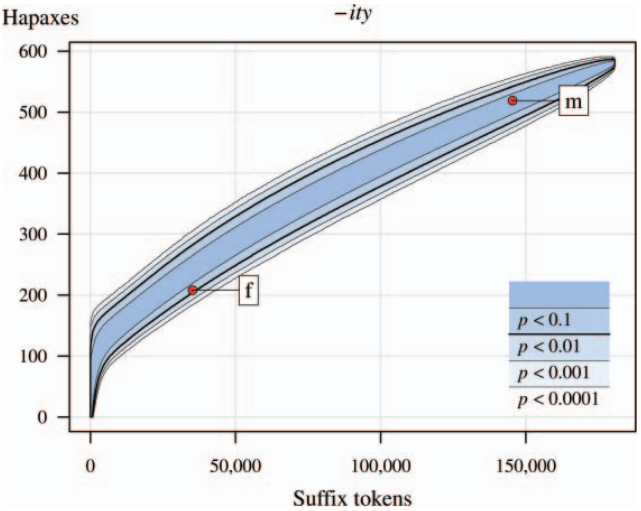

Figure 12. *Gender and* -ity *hapaxes as a function of token frequency in BNC-W_{inf}*

amount of data. In other words, it is no longer quite so much a matter of chance how many hapaxes occur in a subcorpus of a given size. However, it is clear that the number of hapaxes does not grow linearly with either the number of running words (Figure 11) or the number of suffix tokens (Figure 12). This will be discussed further in the next section.

6. Discussion

6.1 *Implications of the choice of productivity measure on sociolinguistic results*

As is evident from Sections 5.1 and 5.2 (as well as Säily and Suomela 2009), type accumulation curves seem to yield quite convincing results with sociolinguistically defined subcorpora; and, as discussed in Section 3, type frequency is a generally accepted measure of (one facet of) morphological productivity. In comparison, hapax accumulation curves seem to yield similar results (see Section 5.3), only less significant; in addition, their use requires a very large corpus, which may not be available for, e.g., earlier varieties of English. Therefore, while the use of hapax accumulation curves in a sociolinguistic approach to morphological productivity may be valid, in practice it is often more feasible and equally valid to use type accumulation instead.

It is, of course, possible that with some other affixes or in another data set, hapax accumulation curves would not yield similar results to type accumulation. Recall that type frequency is seen to reflect *realised* productivity, while hapax-based productivity measures predict the probability of encountering *new* formations (e.g., Baayen 2008). Indeed, Baayen (1993: 182–183) finds that affixes may rank quite differently depending on which measure is used. It is conceivable that, for instance, men and women might use an affix equally diversely, but that men would still be more likely to produce new words with it. With large corpora, then, both type and hapax accumulation curves should be used to ensure a well-rounded picture of productivity.

A potential problem with the accumulation curves approach is that it can control for corpus size in either running words or suffix tokens, but not both at the same time. As noted in Section 3, the results in Sections 5.1 and 5.2 use type frequency as a function of the number of running words in the corpus. It would also have been possible to plot type frequency as a function of token frequency, as in Figure 13, which corresponds to Figure 9 above. Again, the results seem similar but less significant. In future work, it would be of interest to develop the approach further, complementing it with three-dimensional visualisation techniques that would take both measures of corpus size into account simultaneously.[3]

While accumulation curves have proved a promising tool for analysing sociolinguistic variation in productivity, the assessment of Baayen's *P* measure is a more complex question. Recall that *P* corresponds to a point on a hapax accumulation curve with the number of suffix tokens on the *x* axis (see Figure 12). If we wish to compare the *P* figures for men and women, we will immediately run into a problem since the figures are dependent on the size of the subcorpus, and there is much more data from men than from women. In

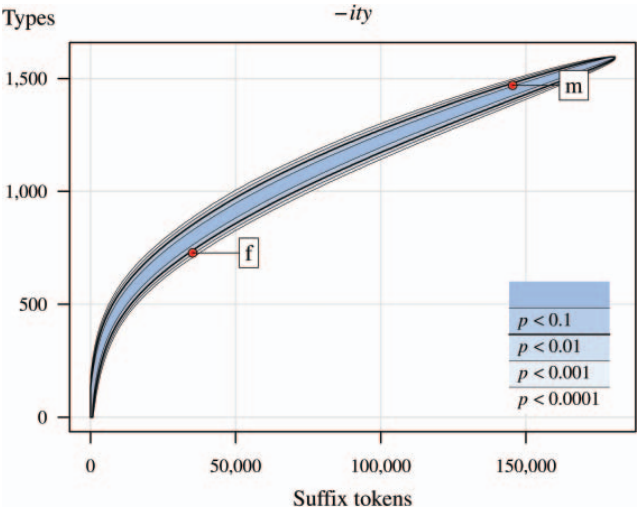

Figure 13. *Gender and* -ity *types as a function of token frequency in BNC-W_{inf}*

essence, *P* assumes that the number of hapaxes grows linearly with the number of suffix tokens ($P = n_1/N$), while Figure 12 clearly shows that it does not do so. This makes *P* figures non-comparable unless the numbers of tokens are of a similar magnitude.

The problem with comparing *P* figures has been noted by Baayen (e.g., 1993: 191), but he does not seem to consider it a serious issue, and continues to recommend the measure as a useful diagnostic (e.g., Baayen 2008). Recall that the theoretical way of looking at *P* is as the tangent to the endpoint of the type accumulation curve. When the comparison is between different suffixes, as in Baayen's work, this may well be feasible. The division by *N* means that suffixes with higher token frequencies are "punished" because a great number of tokens implies the recurrent use of a small number of types – i.e., lower productivity (cf. Baayen 1992: 117).

In sociolinguistic research of the kind described in the present work, however, it makes little sense to "punish" one social group for having a greater number of suffix tokens than another. The number of tokens is dependent on the size of the subcorpus, and if there happens to be four times as much data from men as from women in the corpus, it is not the men's fault, nor does it necessarily reflect the amount of language produced by men as opposed to women in the real world (cf. Keune et al. 2006: 574–575). Hence, I would recommend against using *P* in sociolinguistic studies of this kind. The sociolinguistic usefulness of the measure *P**, which divides hapax frequency by the number of all hapaxes in the (sub)corpus, requires further study.

6.2 *Contributions to corpus-based sociolinguistic inquiry*

This work is one of the first to apply a sociolinguistic approach to the quantitative study of morphological productivity. Together with studies such as Säily and Suomela (2009), it has demonstrated that there is sociolinguistic variation in morphological productivity, that it is measurable using some of the corpus-linguistic methods developed for studying productivity, and that gender emerges as a robust social variable in English derivational morphology, even over time. The last point is just one more piece of evidence for the key role played by women in language change (e.g., Holmes 1999 [1997]; Labov 2001: 262; Nevalainen and Raumolin-Brunberg 2003; Tagliamonte and D'Arcy 2009).

The results of this macro-level study also lend support to the notion of gendered discourse styles (e.g., Holmes 1998; Rayson et al. 1997; Argamon et al. 2003; Säily et al. forthcoming). While the overall picture is that men are more "nouny" and women more "pronouny", this study has shown that the type of noun matters: if it suits their involved style, women may use certain kinds of nouns just as diversely as men, as in the case of *-ness* words in BNC-W. A comparison with the 17th-century results in Säily and Suomela (2009) suggests that this may have been the case for hundreds of years. It is, however, possible that the apparently similar results could stem from different causes. For instance, it is possible that in present-day English, the productivity of *-ness* may be increasing (cf. Baayen and Renouf 1996) and that of *-ity* decreasing. If women are leading the change, this would explain why their use of *-ness* is on a par with men's.[4] The Modern English results may require a different interpretation. Extending the 17th-century study into the long 18th century using the *Corpus of Early English Correspondence Extension* (CEECE) could shed more light on this.

Finally, this study has once again demonstrated the importance of taking into account interaction across categories: in the use of *-ness* in BNC-DS, the influence of gender is tied to social class. It would have been useful if BNC-W had included information on social class as well, and social class is certainly one of the categories that merit closer attention in future studies of gendered discourse styles.

7. Conclusion

The purpose of this work, motivated by a study of 17th-century data by Säily and Suomela (2009), has been twofold. The first aim was to find out whether there is gender-based variation in the productivity of the nominal suffixes *-ness* and *-ity* in present-day British English. This turned out to be the case. Similarly

to the 17th century, women use *-ity* less productively in their writing than men. Considering that women have been shown to use fewer nouns than men in general, the question is why women use *-ness* as diversely as men. A tentative answer is that the semantics of *-ness* could be seen as compatible with the involved discourse style of women – this could be explored further in a more qualitative study. In the spoken subcorpus, however, lower-class women use both *-ness* and *-ity* less diversely than men. This implies a style based on both gender and socio-economic status, on which more research is needed.

The second aim of this work was to analyse the validity of hapax-based measures of morphological productivity in sociolinguistic research of this kind. It was discovered that hapax-based productivity measures require a larger corpus than type-based ones, and that the measure *P* is unusable when comparing subcorpora based on social groups. Otherwise, hapax legomena remain a theoretically well-founded component of productivity measures. It is interesting that the sociolinguistic results obtained using hapax accumulation curves appear similar to but less significant than those obtained using type accumulation curves. This, too, calls for further corpus-linguistic research.

Bionote

Tanja Säily is a PhD student in the Department of Modern Languages at the University of Helsinki. Her thesis is entitled *Sociolinguistic Variation in Derivational Morphology: Studies and Methods in Diachronic Corpus Linguistics*. She is a member of the Research Unit for Variation, Contacts and Change in English (VARIENG) and of the multidisciplinary DAMMOC project, which develops tools and methods for corpus linguistics. E-mail: tanja.saily@helsinki.fi

Notes

* I am grateful to Terttu Nevalainen and Jukka Suomela for discussions and assistance. Thanks also to Harald Baayen and the audience at AACL 2009 for comments on an earlier version of this paper, and to anonymous reviewers and the editors of this issue for helpful feedback. This research was supported in part by Langnet, the Finnish Graduate School in Language Studies.

1. An anonymous reviewer notes that the situation might have been avoided if *-ness* and *-ity* had been treated as variants of a single variable. As mentioned in Section 1, however, this would not have been feasible, since there were too few contexts in which both could have occurred.
2. References to the BNC are given in the format "Subcorpus: Textname S-unit number"; emphases are mine.
3. Thanks to an anonymous reviewer for this suggestion.
4. Thanks to an anonymous reviewer for drawing my attention to this possibility.

References

Adamson, Sylvia. 1989. With double tongue: Diglossia, stylistics and the teaching of English. In Michael Short (ed.), *Reading, analysing and teaching literature*, 204–240. London: Longman.

Anshen, Frank & Mark Aronoff. 1989. Morphological productivity, word frequency and the Oxford English Dictionary. In Ralph W. Fasold and Deborah Schiffrin (eds.), *Language change and variation*, 197–202. Amsterdam: John Benjamins.

Argamon, Shlomo, Moshe Koppel, Jonathan Fine & Anat Rachel Shimoni. 2003. Gender, genre, and writing style in formal written texts. *Text* 23(3). 321–346.

Aronoff, Mark. 1976. *Word formation in generative grammar* (Linguistic Inquiry Monograph One). Cambridge, MA: The MIT Press.

Baayen, R. H. 1992. Quantitative aspects of morphological productivity. In Geert Booij & Jaap van Marle (eds.), *Yearbook of morphology 1991*, 109–149. Dordrecht: Kluwer Academic Publishers.

Baayen, R. H. 1993. On frequency, transparency and productivity. In Geert Booij & Jaap van Marle (eds.), *Yearbook of morphology 1992*, 181–208. Dordrecht: Kluwer Academic Publishers.

Baayen, R. H. 2008. Corpus linguistics in morphology: Morphological productivity. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics: An international handbook*, 899–919. Berlin: Mouton de Gruyter.

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

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

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, Douglas & Jená Burges. 2000. Historical change in the language use of women and men: Gender differences in dramatic dialogue. *Journal of English Linguistics* 28(1). 21–37.

Biber, Douglas & Edward Finegan. 1989. Drift and the evolution of English style: A history of three genres. *Language* 65(3). 487–517.

BNC = *The British National Corpus*, version 3 (BNC XML Edition). 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. http://www.natcorp.ox.ac.uk/ (accessed 21 January 2010).

Bolinger, Dwight L. 1948. On defining the morpheme. *Word* 4. 18–23.

Burnard, Lou (ed.). 2007. *Reference guide for the British National Corpus (XML edition)*. Published for the British National Corpus Consortium by the Research Technologies Service at Oxford University Computing Services. http://www.natcorp.ox.ac.uk/docs/URG/ (accessed 21 January 2010).

Cameron, Deborah. 2006. Gender. In Keith Brown (ed.), *Encyclopedia of language and linguistics*, 733–739. Oxford: Elsevier. DOI: 10.1016/B0-08-044854-2/01463-2 (accessed 19 January 2010).

Cameron, Deborah. 2008. Issues of gender in modern English. In Haruko Momma & Michael Matto (eds.), *A companion to the history of the English language*, 292–302. Chichester: Wiley-Blackwell.

CEEC = *Corpus of Early English Correspondence*. 1998. Compiled by Terttu Nevalainen, Helena Raumolin-Brunberg, Jukka Keränen, Minna Nevala, Arja Nurmi & Minna Palander-Collin at the Department of English, University of Helsinki.

CEECE = *Corpus of Early English Correspondence Extension*. Compiled by Samuli Kaislaniemi, Mikko Laitinen, Minna Nevala, Terttu Nevalainen, Arja Nurmi, Minna Palander-Collin, Helena Raumolin-Brunberg & Anni Sairio at the Department of English, University of Helsinki.

Cowie, Claire & Christiane Dalton-Puffer. 2002. Diachronic word-formation and studying changes in productivity over time: Theoretical and methodological considerations. In Javier E. Díaz Vera (ed.), *A changing world of words: Studies in English historical lexicography, lexicology and semantics*, 410–437. Amsterdam: Rodopi.

Dalton-Puffer, Christiane. 1996. *The French influence on Middle English morphology: A corpus-based study of derivation*. Berlin: Mouton de Gruyter.

Gaeta, Livio & Davide Ricca. 2006. Productivity in Italian word formation: A variable-corpus approach. *Linguistics* 44(1). 57–89.

Hay, Jennifer & R. H. Baayen. 2003. Phonotactics, parsing and productivity. *Italian Journal of Linguistics* 15(1). 99–130.

Hoffmann, Sebastian, Stefan Evert, Nicholas Smith, David Lee & Ylva Berglund Prytz. 2008. *Corpus linguistics with* BNCweb – *a practical guide* (English Corpus Linguistics 6). Frankfurt am Main: Peter Lang.

Holmes, Janet. 1998. Women's talk: The question of sociolinguistic universals. In Jennifer Coates (ed.), *Language and gender: A reader*, 461–483. Malden, MA & Oxford: Blackwell.

Holmes, Janet. 1999 [1997]. Setting new standards: Sound changes and gender in New Zealand English. *English World-Wide* 18(1). 107–142. Reprinted in Juan Camilo Conde-Silvestre & Juan Manuel Hernandez-Campoy (eds.), Variation and linguistic change in English: Diachronic and synchronic studies. [Special issue]. *Cuadernos de Filología Inglesa* 8. 147–175.

Holmes, Janet & Miriam Meyerhoff. 2003. Different voices, different views: An introduction to current research in language and gender. In Janet Holmes & Miriam Meyerhoff (eds.), *The handbook of language and gender*, 1–17. Oxford: Blackwell.

Hudson, R. A. 1996. *Sociolinguistics*, 2nd edn. Cambridge: Cambridge University Press.

Kastovsky, Dieter. 1986. The problem of productivity in word formation. *Linguistics: An Interdisciplinary Journal of the Language Sciences* 24(3[283]). 585–600.

Keune, Karen, Roeland van Hout & R. H. Baayen. 2006. Socio-geographic variation in morphological productivity in spoken Dutch: A comparison of statistical techniques. In Jean-Marie Viprey (ed.), *Proceedings of the 8th international conference on the statistical analysis of textual data*, vol. 2, 571–581. Besançon: Presses Universitaires de Franche-Comte.

Labov, William. 2001. *Principles of linguistic change, volume 2: Social factors*. Malden, MA & Oxford: Blackwell.

Marchand, Hans. 1969. *The categories and types of present-day English word-formation: A synchronic-diachronic approach*, 2nd edn. Munich: C. H. Beck'sche Verlagsbuchhandlung.

Milroy, Lesley & Matthew Gordon. 2003. *Sociolinguistics: Method and interpretation* (Language in Society 34). Malden, MA: Blackwell.

Nevalainen, Terttu. 2006. Synchronic and diachronic variation. In Keith Brown (ed.), *Encyclopedia of language and linguistics*, 356–363. Oxford: Elsevier. DOI: 10.1016/B0-08-044854-2/01521-2 (accessed 21 January 2010).

Nevalainen, Terttu & Helena Raumolin-Brunberg. 2003. *Historical sociolinguistics: Language change in Tudor and Stuart England* (Longman Linguistics Library). London: Pearson Education.

Nurmi, Arja, Minna Nevala & Minna Palander-Collin (eds.). 2009. *The language of daily life in England (1400–1800)* (Pragmatics & Beyond New Series 183). Amsterdam: John Benjamins.

Palander-Collin, Minna. 1999. *Grammaticalization and social embedding: I THINK and ME-THINKS in Middle and Early Modern English* (Mémoires de la Société Néophilologique 55). Helsinki: Société Néophilologique.

Plag, Ingo, Christiane Dalton-Puffer & R. H. Baayen. 1999. Morphological productivity across speech and writing. *English Language and Linguistics* 3(2). 209–228.

Rayson, Paul, Geoffrey Leech & Mary Hodges. 1997. Social differentiation in the use of English vocabulary: Some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics* 2(1). 133–152.

Riddle, Elizabeth M. 1985. A historical perspective on the productivity of the suffixes *-ness* and *-ity*. In Jacek Fisiak (ed.), *Historical semantics; historical word-formation*, 435–461. Berlin: Mouton de Gruyter.

Romaine, Suzanne. 1985. Variability in word formation patterns and productivity in the history of English. In Jacek Fisiak (ed.), *Papers from the 6th international conference on historical linguistics*, 451–465. Amsterdam: John Benjamins.

Säily, Tanja. 2008. *Productivity of the suffixes* -ness *and* -ity *in 17th-century English letters: A sociolinguistic approach*. Helsinki: University of Helsinki MA thesis. http://urn.fi/URN:NBN:fi-fe200810081995 (accessed 21 January 2010).

Säily, Tanja & Jukka Suomela. 2009. Comparing type counts: The case of women, men and *-ity* in early English letters. In Antoinette Renouf & Andrew Kehoe (eds.), *Corpus linguistics: Refinements and reassessments* (Language and Computers: Studies in Practical Linguistics 69), 87–109. Amsterdam: Rodopi.

Säily, Tanja, Terttu Nevalainen & Harri Siirtola. Forthcoming. Variation in noun and pronoun frequencies in a sociohistorical corpus of English.

Suomela, Jukka. 2007. Type and hapax accumulation curves. Computer program. http://www.cs.helsinki.fi/jukka.suomela/types/ (accessed 21 January 2010).

Tagliamonte, Sali A. & Alexandra D'Arcy. 2007. Frequency and variation in the community grammar: Tracking a new change through the generations. *Language Variation and Change* 19. 199–217.

Tagliamonte, Sali A. & Alexandra D'Arcy. 2009. Peaks beyond phonology: Adolescence, incrementation, and language change. *Language* 85(1). 58–108.

Talbot, Mary. 2006. Gender and language. In Keith Brown (ed.), *Encyclopedia of language and linguistics*, 740–742. Oxford: Elsevier. DOI: 10.1016/B0-08-044854-2/00331-X (accessed 21 January 2010).