Derivational Productivity and Text Typology

R. Harald Baayen[*]
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
e-mail baayen@mpi.nl

to appear in **Journal of Quantitative Linguistics**

Abstract

The productivity of English derivational affixes is studied as a function of text type. Principal component analyses show that texts can be adequately classified not only on the basis of the relative frequencies of the highest frequency words (Burrows, 1992, 1993), but also on the basis of the productivity of derivational affixes. Stylistically heterogeneous texts are clustered into text types, stylistically homogeneous texts cluster in the time dimension, allowing diachronic changes in productivity to be traced. Supplementary analyses on the basis of the relative frequencies of function words support the morphology-based clusterings. The role and marked nature of the nonnative stratum of the lexicon is discussed in detail, as well as the way in which the rival affixes *-ness* and *-ity*, and *un-* and *in-*, are put to use. The results obtained show that any theory of morphological productivity that does not take stylistic factors into account is incomplete.

Keywords: derivational morphology, productivity, text typology, markedness, function words.

1 Introduction

Most of the research on morphological productivity, the possibility for speakers of a language to effortlessly use and understand novel rule-based polymorphemic words, has focussed predominantly on the formal and semantic properties a word formation rule should have for it to be productive (Aronoff, 1976, Booij, 1977, Rainer, 1988). Although some researchers have suggested that speech register and text type may co-determine the productivity of a word formation rule (Burgschmidt, 1977; Romaine, 1983), this possibility has never been investigated in detail. In fact, linguists such as Biber (Biber, 1989, Biber and Finegan, 1989) working in the area of genre-oriented text typology and literary scholars such as Burrows (Burrows, 1992, 1993) do not take derivational morphology into account in their multivariate analyses. Nevertheless, the way in which the morphological resources of a language are exploited may well be co-determined by sociolinguistic and stylistic factors. A first aim of the present pilot study is to investigate whether the way in which authors put the affixes of their language to use in written texts can serve as a basis for establishing a text typology. Since a host of syntactic, semantic and pragmatic factors (Biber, 1989) are also, and perhaps even more strongly, correlated with text type, a morphology-based text typology on its own will be less accurate than a comprehensive analysis in which syntactic, semantic and pragmatic factors are considered jointly. However, if we find that reasonable text classifications can be obtained on the basis of the use of derivational affixes only, it seems likely that the accuracy of more general typological analyses may be increased by also taking morphological data into account. A second aim of this paper is to gain some insight into the extent to which

[*] I am indebted to Geert Booij, John Burrows, Shelly Lieber, Christian Mair, Roswitha Raab-Fischer, and Rob Schreuder for valuable discussion.

the productivity of word formation rules is influenced by text type on the one hand and author specific idiosyncracies on the other. A third aim is to trace possible positive and negative correlations between the use of word formation rules.

Our discussion is structured as follows. Section 2 is a brief introduction to the concept of morphological productivity. Section 3 discusses the methodology underlying our analyses: a principle components analysis on the basis of the values observed for a wide range of affixes of a quantitative measure for the degree of productivity. Section 4 shows that this method leads to a reasonable text classification. Section 5 shows that, when applied to the relative frequencies of the most frequent words (Burrows 1992, 1993), a principal components analysis gives rise to similar groupings. In section 6 the same methodology is applied to the novels in our sample to study effects of morphology within a single text type. Section 7 discusses the theoretical implications of the results obtained for the study of morphological productivity on the one hand and literary analysis on the other.

2 Productivity

Word formation rules generally differ with respect to how often they are used for producing or understanding novel forms. Some rules are quite productive in that they give rise to large numbers of neologisms. Other rules appear to be descriptive only, in the sense that they describe the structure of existing complex words without giving rise to new formations. Most studies on morphological productivity have focussed on how phonological, syntactic and semantic restrictions on affixation constrain the set of possible words with a given affix. These qualitative studies generally take the degree of productivity to be inversely proportional to the number of restrictions (such as the restriction barring comparative *-er* from attaching to polysyllabic adjectives) that define a rule's input domain (Booij, 1977). Unfortunately, this qualitative definition cannot serve as the basis for an operational quantitative definition of the notion 'degree of productivity' (see Baayen, 1989, 1992, 1993, Baayen and Lieber, 1991). In the light of the widely varying numbers of different words (or word types, as opposed to word tokens) in which particular affixes appear in a single text corpus, the general usefulness of a word formation rule should be acknowledged as a factor co-determining its productivity. In Dutch, for example, pejorative personal names in *-erd* (*gekkerd*, 'fool', from *gek*, 'foolish') seldom appear in written texts, in contrast to personal (agent) nouns in *-er* (*gever*, 'giver'). But even in colloquial conversation, where formations in *-erd* are more productive, this suffix does not give rise to large numbers of formations: there are cultural limits to its use. The substantial difference between the degree of productivity of *-erd* and that of *-er* cannot be explained in terms of structural factors only. As pointed out by Van Santen (1992), degrees of productivity should not be coupled with the number of restrictions defining the input domain of a word formation rule. Instead, the study of degrees of productivity should focus on the variability that characterizes the extent to which rules are applied to the words satisfying their input constraints. Although van Santen does not discuss the possibility that stylistic factors may be strong determinants of the potentiality of word formation for given input domains, her theoretical position provides a fruitful starting point for the study of the role of social and stylistic factors.

Very little is known about the nature and strength of non-structural factors. There are numerous hints in the literature that the use of affixes is influenced by such factors. For instance, van Haeringen (1971), in a detailed study of the Dutch nominalizing suffix *-ing*, intuitively judged this affix to be more productive in more formal Dutch. Burgschmidt (1977) explicitly incorporates the appropriateness of a particular kind of word formation pattern for a given speech register as a factor co-determining productivity in his theory of word formation. Unfortunately, the examples he adduces in support of his claim that speech register co-determines productivity are of an anecdotal nature only. A somewhat more detailed study of the possible role of sociolinguistic factors can be found in Romaine (1983). She discusses an experiment in which subjects were asked to judge the acceptability of attaching the suffixes *-ness* and *-ity* to 100 different adjectives. The resulting judgements suggest that the acceptability of *-ness*, of *-ity*, or of both suffixes for a given base word is correlated with the age of the speaker.[1] This is an important finding, but Romaine's experimental method

[1] A loglinear analysis of Romaine's data shows that sex does not guide affix choice, contrary to what Romaine (1983:187–188)

precludes the possibility of unearthing the possible role of the stylistic factors that normally co-determine affix choice. The present paper is an attempt to develop a methodology by means of which the role of various 'sociolinguistic' as well as stylistic factors can be studied in more detail.

3 Methodology

In order to study differences in the way various authors exploit the word formation rules of their language, we first need a quantitative formalization of the notion 'degree of productivity'. The formalization that we will use here defines the degree of productivity of a word formation rule in terms of its relative contribution to the growth rate of the vocabulary. Consider figure 1, which summarizes how the vocabulary of E. Bronte's *Wuthering Heights* develops through 'text time' t for $t = 0, 250, 500, \ldots, 25000$ word tokens. The upper left hand panel shows how the vocabulary size increases as a function of the text (or sample) size t. The upper right hand panel plots the rate at which the vocabulary increases. This growth rate, which can be expressed mathematically as $\mathrm{E}[V_t(1)]/t$, where E is the expectation operator and $V_t(1)$ denotes the number of types observed once only (the so-called hapax legomena) among the t tokens of the current text size, is a decreasing

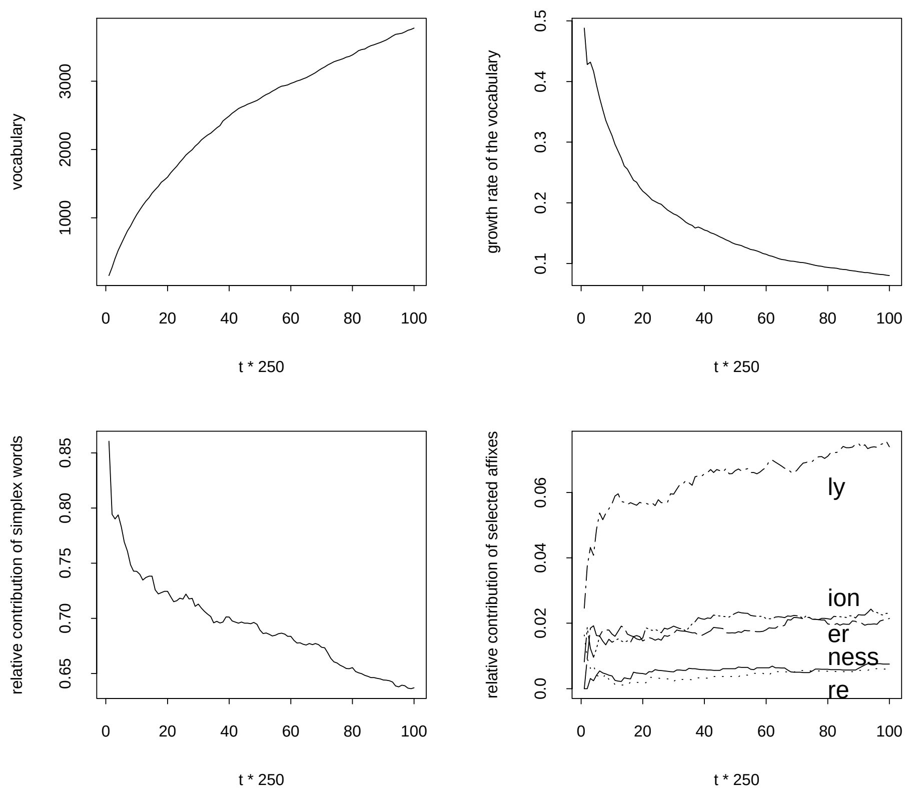

Figure 1: Growth curve of the vocabulary, the growth rate of the vocabulary, the contribution of monomorphemic words to the growth rate, and the contribution of selected affixes to the growth rate, as a function of the 'text time' t, for E.Bronte's *Wuthering Heights.* The curves are based on 100 measurements taken at intervals of 250 word tokens.

tentatively suggests.

function of t (for technical details and further references to the literature on the statistics of word frequency distributions the reader is referred to Baayen (1993b) and Chitashvili and Baayen (in press)). Words of different morphological constituencies contribute to this growth rate. The lower left hand panel shows that although initially monomorphemic words are predominant among the hapaxes, the relative contribution of such words to the growth rate of the vocabulary decreases with t. Conversely, as shown in the lower right hand panel of figure 1, the relative contributions of productive affixes are (slowly) increasing functions of the text size. Thus it seems natural to gauge the degree of productivity $\mathcal{P}_a^*$ of an affix a in terms of its relative contribution to the growth rate of the vocabulary (Baayen, 1993):

$$\mathcal{P}_a^* = \frac{\mathrm{E}[V_{a,t}(1)]/t}{\mathrm{E}[V_t(1)]/t} = \frac{\mathrm{E}[V_{a,t}(1)]}{\mathrm{E}[V_t(1)]}. \tag{1}$$

As t increases, the accuracy with which we estimate the likelihood of encountering neologisms (and very low-frequency complex words that require rule-based processing in the absence of strong enough memory traces) also increases (see Baayen, 1992, 1993). For the smallish texts underlying our analyses ($t = 25,000$), many morphologically complex hapaxes are well-known and well-established English words. Even at this small sample size, however, rough estimates of the relative sizes of the growth rates of a set of affixes can already be obtained.[2] Note that in this approach the degree of productivity of some affix is viewed as the outcome of linguistic structural factors and various social factors that jointly determine the statistical readiness with which a rule is put to use.

Given this quantitative productivity measure, we should now consider what texts to choose for our analysis. Two considerations are relevant here. First, a wide variety of texts should be selected. Second, since the productivity measure $\mathcal{P}^*$ becomes more accurate as t increases, the texts should not be too small. Forced by practical considerations, the texts should therefore be available in electronic form. Fortunately, a reasonable variety of electronic texts is available by anonymous ftp. From the Online Book Initiative (OBI) at obi.std.com, the Project Gutenberg (PG) at mrcnext.cso.uiuc.edu, and the Oxford Text Archive (OTA) at black.ox.ac.uk, a total of 44 texts ranging from children's books to officialese and from well-known literary texts to a Startrek novel were selected for analysis. The selected texts are documented in the appendix. Most of these texts are by nineteenth or early twentieth century novelists. In order to obtain a wide enough variety of texts, some earlier texts (Luke–Acts in the King James Version, Milton's *Paradise Lost*, Jane Austen's *Pride and Prejudice*) and some modern texts (officialese such as documents from the U.S. Accounting Office) were also included. Obviously, the present sample is not ideal in that a number of texts included are not prototypical instances of the provisional text types considered in this study. Conversely, prototypical examples of, for instance, contemporary children's literature are not included. Hence differences between text types emerging from the present study may well be less substantial than those one may expect for the analysis of a more representative sample.

Since the growth rate of the vocabulary is a function of text size, the same number of word tokens has to be analyzed for each text. In this study the first 25,000 words were selected — roughly the size of the smallest complete novel in our sample. For each text, we obtained the morphological structure of its constituent words by means of PC-KIMMO, a parser developed by Antworth (fully documented in Antworth (1990) and freely available by anonymous ftp from the Consortium for Lexical Research at clr.nmsu.edu). Where necessary, the analyses of PC-KIMMO were post-edited by hand. For each of the 44 texts studied here, we calculated the $\mathcal{P}^*$ productivity statistic for the derivational affixes *-ness, -ity, -ment,* agentive *-er, -ee, -ism, -ian, -ation, -able, -ful, -y,* comparative *-er,* superlative *-est, -less,* adjectivizing *un-* and *in-*, verb-forming *un-*, *-ize, -ify, de-, re-, en-,* adverbial *-ly, ex-, anti-, semi-*, and the intensifying prefixes *mega-, hyper-, ultra-*, and *super-*. The summed relative contribution R of these affixes to the growth rate of the vocabulary of a particular text,

$$R = \sum_a \mathcal{P}_a^* = \frac{\sum_a \mathrm{E}[V_{a,t}(1)]}{\mathrm{E}[V_t(1)]}. \tag{2}$$

[2] Note that by using $\mathcal{P}^*$ rather than V (an unreliable productivity measure, especially for small t), we avoid the problem how to weight types for their frequency of occurrence: all types figuring in the analysis occur once only, no weighting is necessary.

was also obtained.

As a first step, we may consider whether global differences in the use of morphology can be observed for our selection of texts by plotting the texts in the plane defined by the vocabulary size V and the summed relative contribution to the growth rate R, as shown in figure 2.[3] Adopting a provisory classification into Children's books (coded by an initial C in the abbreviations used in the figures below; for further details the reader is referred to the appendix), Literary texts (coded by an initial L), Officialese and scientific texts (coded by an initial O) and Religious Texts (coded by an initial B), we find that this classification is to some extent reflected in the values of R and V. The literary texts appear in the larger upper right hand corner.

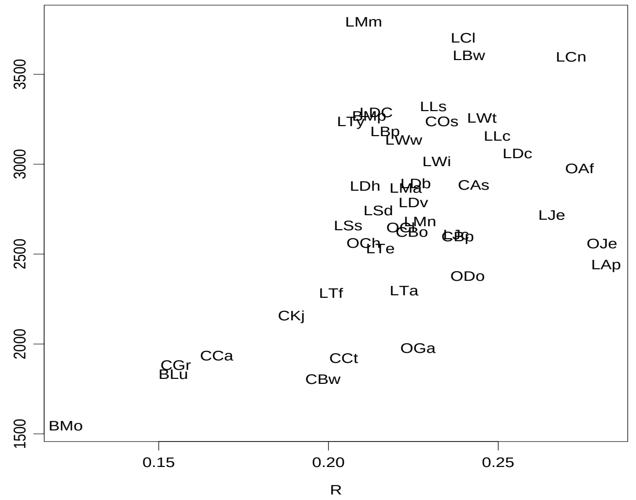

Figure 2: Vocabulary size V and relative contribution of morphology R for 44 texts of size $t = 25,000$.

They are characterized by the more substantial vocabulary sizes as well as by the larger R-values. The officialese tends to appear with smaller V and perhaps slightly higher values of R, but these texts cannot be said to cluster as a distinct group. The majority of the children's books form a cluster in the lower left hand corner. The religious texts occupy an even more extreme position: they are characterized by extremely low values for both R and V. Note that there is no simple linear relation between R and V, that is, it is not possible to predict V given R or vice versa. Some authors, Melville (LMm) for instance, make relatively little use of word formation. In the case of Melville, this is compensated for by an extensive use of monomorphemic words and (synchronically unproductive) complex words. Conversely, authors such as W. James (OJe) and Austen (LAp) make extensive use of word formation, and relatively little use of non-derived words.

Having observed that a very rough classification into three text types (religious texts versus children's books versus novels and officialese) can already be obtained on the basis of the simple statistics V and R, we now turn to consider in detail how authors put individual affixes to use. To do so, we make use of a principal components analysis. Our collection of texts constitutes a sample with 44 multivariate observations. Each observation has 27 'responses' or dimensions, one for each affix.[4] Thus we have 44 points in a 27-dimensional 'affix space'. Rather than attempting to study the relations between the 44 texts in this multi-dimensional space as such, we use a principal components analysis to reduce the number of dimensions. Such an analysis allows us to extract from our 27-dimensional space those components (or new dimensions) that account for the major part of the variance. Similar texts will appear in roughly the same region of the space spanned

[3] The non-standard spellings in Milton's *Paradise Lost* are not analyzed by PC-KIMMO. Such words appear as independent word types in our counts, giving rise to a somewhat inflated value for the vocabulary size V.

[4] The intensifying prefixes are considered as a group. They are referred to by their most productive member, *super*.

by these principal components. Moreover, by studying which affixes are most closely correlated with the principal components, an interpretation of the dimensions of the reduced 'affix space' may be obtained.

4 Affixes and Text Types

The results of applying a principal components analysis to the covariance matrix[5] of the data matrix $(\mathcal{P}^*_{c,a}), c = 1, 2, \ldots, 44, a = 1, 2, \ldots, 27$ are summarized in figures 3 and 4. The first three components

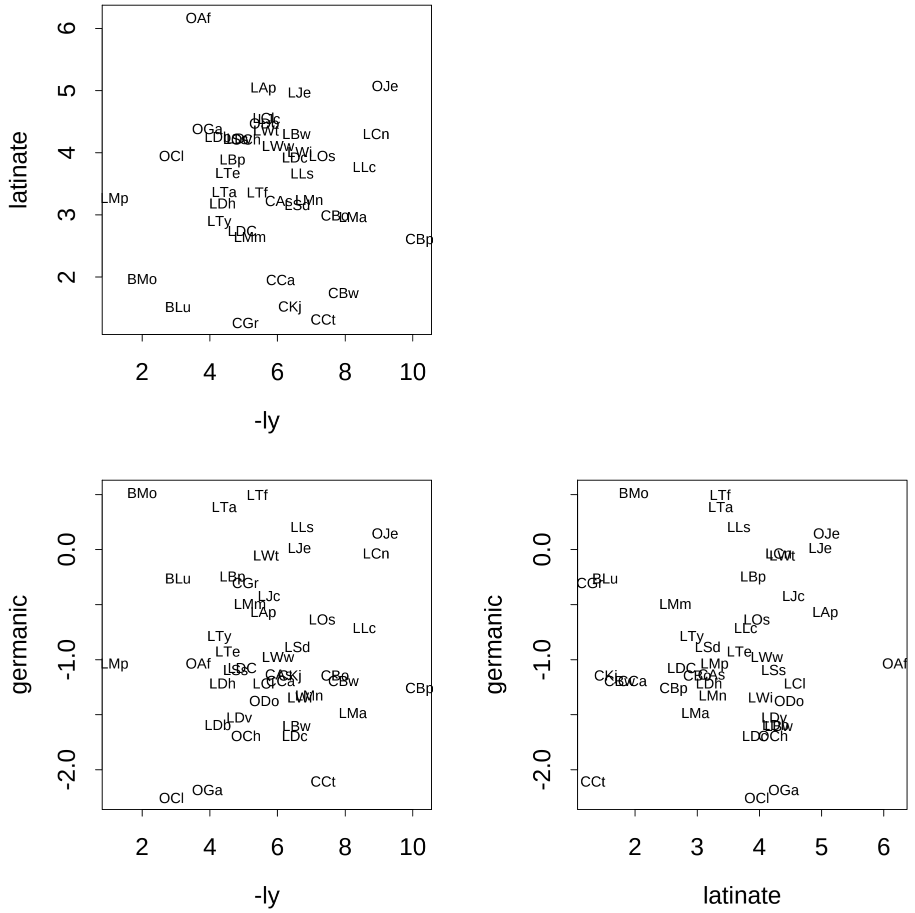

Figure 3: Morphology-based principal components analysis for 44 texts. The scatterplots chart the three-dimensional space spanned by the three significant principal components. The three panels can be viewed as the top, the front and the right-hand side of a transparent cube in which the texts are located. The first letter of the codes denote the text type. For further details see the appendix.

account for some 75% of the variance. The first component explains 51.2% of the variance, the second

[5] The analysis is carried out on the covariance matrix rather than on the correlation matrix. All variables considered here are reasonably commensurable, in which case the covariance matrix is to be preferred (cf. Morrison, 1976:268). The use of the correlation matrix, implying standardization of all measurements, would also have the undesirable effect of obscuring differences in degrees of productivity.

18.1%, and the third 7.3%. The first component is fully correlated with the suffix *-ly* ($r_{\text{-ly},1} = 1.00$). No other affixes have a large ($|r| > 0.4$) correlation coefficient for this component. The second and third components represent affixes from the latinate and germanic strata of the lexicon respectively. The latinate affixes *-ation* ($r_{\text{-ation},2} = 0.90$), *in-* ($r_{\text{in-},2} = 0.78$), *-ity* ($r_{\text{-ity},2} = 0.76$), and *-ment* ($r_{\text{-ment},2} = 0.53$) are strongly correlated with the second dimension. The adjectivizing prefix *un-* ($r_{\text{un-},2} = 0.66$) is the only germanic affix with a strong positive correlation. Other germanic affixes such as *-est* and comparative *-er* show up with negative correlation coefficients ($r_{\text{-est},2} = -0.36$, $r_{\text{-er},2} = -0.36$). Pending further discussion in section 7, I will refer to this dimension, that represents a scale of nativeness, as the nonnative or latinate principal component. The germanic affixes *-ness* ($r_{\text{-ness},3} = 0.86$), agentive *-er* ($r_{\text{-er (a)},3} = -0.60$) and comparative *-er* ($r_{\text{-er (c)},3} = -0.54$) are correlated with the third dimension. These are the only affixes that show up on this component with a large correlation coefficient ($|r| > 0.4$). We will refer to this component as the native or germanic dimension. Contrary to the second principal component, however, this component does not measure degrees of (non)nativeness — it is sensitive to differences in use within the set of germanic affixes.

Figure 3 presents a scatterplot matrix locating our 44 texts in the three planes defined by these dimensions. Recall that roughly 50% of the variance is due to how authors make use of the adverbializing suffix *-ly*. The general pattern seems to be that the religious texts (BMp, BLu), Milton (LMp), and the officialese and scientific texts (The Federalist Papers (OAf), Clinton's speeches (OCl), the texts from the Government Accounting Office (OGa), the Congress Hearings (OCh), and Darwin's *On the Origin of the Species* (ODo)) tend to use *-ly* sparingly. Conversely, novels, whether written for adults or children, but also W. James' *Essays in Radical Empiricism* (OJe)), exploit *-ly* more fully.

Next consider the second and third dimensions. Children's books score low on the second component. As expected, they show a tendency to avoid the use of latinate morphology. They show a preference for the use of *-est* and comparative *-er*. Conversely, officialese and scientific texts tend to score rather high on the latinate dimension. The majority of the literary novels in our sample are found in the intermediate range, but novels such as Austen's *Pride and Prejudice* (LAp) and Henry James' *The Europeans* (LJe) also reveal an abundant use of latinate affixes. Turning to the third dimension, we find that the officialese of Clinton (OCl) and the Government Accounting Office (OGa) score low on the germanic affixes, indicating a preference for agentive *-er* and a slight tendency to make more use of the prefix *re-* ($r_{\text{re-},2} = 0.37$, $r_{\text{re-},3} = -0.39$). The *Book of Mormon* (BMo), two of Trollope's novels (*Can you forgive her?*, LTf, and *Ayala's Angel*, LTa), as well as London's *The Sea Wolf* (LLs) score high on this dimension, and the same holds for W. James' *Essays in Radical Empiricism* (OJe). These texts show a preference for *-ness*.

Having observed where our texts are positioned on the three significant dimensions, we are now in the position to consider whether these texts form more or less distinct clusters corresponding to our crude typology of literary novels, officialese, children's books and religious texts. To answer this question, we need a three-dimensional scatterplot. Although the scatterplot matrix of figure 3 can be used to build a mental image of such a three-dimensional scatterplot, it is more convenient to use the graphical tools 'brush' and 'spin' of the Splus statistical programming environment (Becker, Chambers and Wilks, 1988, StatSci, 1991). These tools enable one to construct a three-dimensional representation by rotating the cloud of data points along with the axes spanning the three significant dimensions. Figure 4 is a screen dump of the clustering that emerges when these tools are applied to the present data. The axes 1 and 3 point backwards. The literary novels are marked by dots, the children's books by the smallest squares, and the officialese by largest squares. The religious texts (Luke/Acts and Milton) are represented by the intermediately sized squares. What we find is that the four text types occupy reasonably distinct regions in 'morphological space', with the literary novels occupying the central region and the other texts appearing at the periphery. Not surprisingly, the germanic and latinate strata of the lexicon play a major role in teasing apart 'opposite' types such as children's books and officialese. We may conclude that text types can indeed be distinguished on the basis of how productively they exploit the morphological rules of the language.

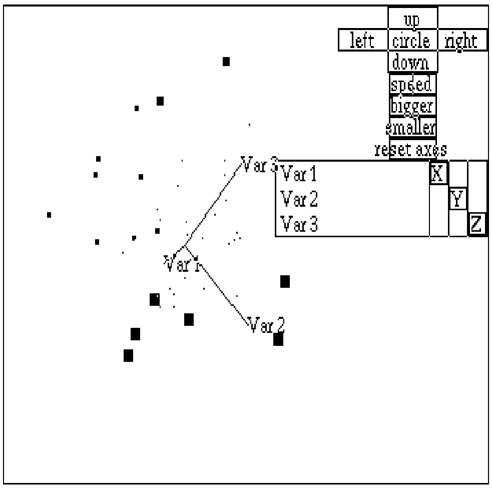

Figure 4: Location of texts in the three-dimensional space defined by the significant principal components of a morphology-based analysis.

5 Function Words and Text Types

The above conclusion would be strengthened if it could be shown that a similar clustering can be obtained independently of the morphological data. Burrows (1992, 1993) discusses results showing that differences between authors with respect to factors such as age or country of origin can be ascertained on the basis of counts of the highest frequency words of the language. Hence it seems worthwhile to investigate whether a clustering of our texts on the basis of the relative frequencies of these words can be obtained that supports the morphology-based clustering. Closely following Burrrows' analysis, we subjected the relative frequencies of the 40 most frequent words (table 1) of the pooled vocabulary of our sample of 44 texts to a principal components analysis. Among these words we find the definite and indefinite articles, various conjunctions, the personal pronouns, the negations *no* and *not*, a number of prepositions, and the verbal forms *be, is, was, are, were, have, had* and *said*. Note that PC-KIMMO does not collapse the irregular verbs. For ease of reference, I will henceforth refer to these highest-frequency words as function words.

the	*and*	*of*	*to*	*a*	*I*	*in*	*that*	*it*	*he*
was	*his*	*you*	*with*	*as*	*for*	*had*	*is*	*but*	*not*
be	*at*	*on*	*they*	*said*	*have*	*all*	*this*	*by*	*which*
me	*from*	*so*	*we*	*were*	*are*	*there*	*or*	*them*	*no*

Table 1: The 40 most frequent words in the pooled vocabulary of the 44 texts listed in the appendix.

The results of a principal components analysis carried out on the correlation matrix of the relative frequencies with which these function words are used are summarized in figure 5 by means of scatterplots for the first three (of six) significant components. The first dimension, which accounts for 23.5% of the variance, is positively correlated with the verbal forms *was* ($r_{\text{was},1} = 0.83$), *had* ($r_{\text{had},1} = 0.70$), and *said* ($r_{\text{said},1} = 0.63$). It is negatively correlated with *are* ($r_{\text{are},1} = -0.65$). This suggests that the first principal component presents a scale of narrativity. In addition, the preposition *by* is negatively correlated with this component ($r_{\text{by},1} = -0.71$), which might indicate differences in the use of the passive voice. Turning to the second principal component (13.4%), we observe positive correlations with the verb forms *have* ($r_{\text{have},1} = 0.78$) and *is* ($r_{\text{is},1} = 0.60$), and the demonstrative/complementizer/relative pronoun *that* ($r_{\text{that},1} = 0.69$). Texts favoring the present tense, possibly with a preference for subordination, score high

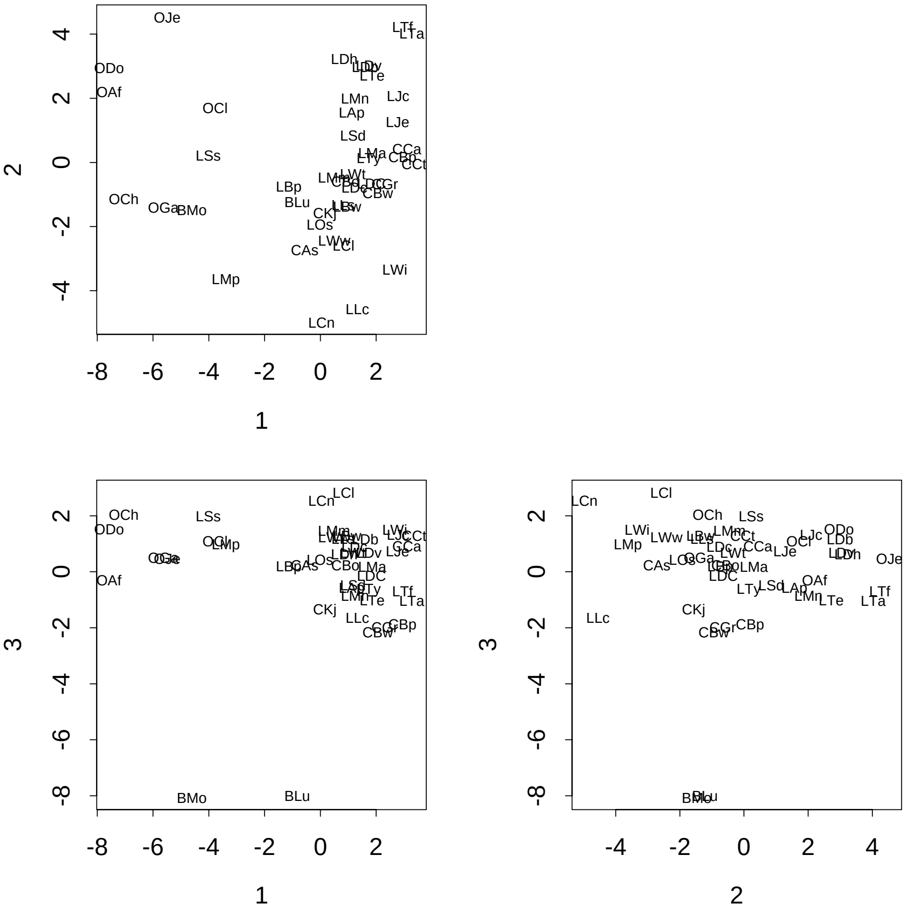

Figure 5: Function word based principal components analysis for 44 texts. The scatterplots chart the three-dimensional space spanned by the first three significant principal components. The three panels can be viewed as the top, the front and the right-hand side of a transparent cube in which the texts are located. The first letter of the codes denote the text type. For further details see the appendix.

on this dimension. The third principal component (11.4%) shows a strong negative correlation with the use of *them* ($r_{\text{them},1} = -0.89$). The same holds for the conjunction *and* ($r_{\text{and},1} = -0.64$). The higher significant principal components account for small proportions of the variance only, and do not reveal interpretable patterns.

Figure 5 shows that the first and third principal components succeed in teasing apart the officialese, the religious texts, and the narrative texts (the children's books and the adult novels jointly). The first dimension separates the narrative from the non-narrative texts. The only exceptions are Chu's *More than a Chance Meeting* and Milton's *Paradise Lost*, which cluster with the officialese. Possibly, Chu's Startrek novel reveals the background of the author who, to judge from the header of his electronic novel, is a computer engineer at the University of Oklahoma. The reason that Milton's *Paradise Lost* behaves exceptionally may reside in Milton's frequent use of the present tense for the many evocative descriptions of general religious truths that are found interleaved with narratives episodes using the past tense. Turning to the third dimension, we find that it singles out the religious texts. These texts are characterized by an intense use of *them* and *and*. The appearance of the latter conjunction is to be expected given the extensive parataxis characteristic of the texts in Biblical Hebrew that have influenced both the Greek author of Luke/Acts and Joseph Smith. Finally note that the second dimension does not appear to have classificatory relevance. We may conclude that, even though the children's books do not separate as well from the adult novels as in the case of a morphology-based analysis, the emerging pattern provides independent support for the morphology-based classification of our texts.

6 Affixes and Function Words in the Novels

Having observed that both affixes and function words can be used to cluster a wide variety of texts into more or less distinct text types, we now turn to consider the question what results might be obtained if a far more homogeneous set of texts is selected for analysis. Burrows' (1992, 1993) studies show that for such samples the principal components may uncover factors such as date of birth, sex, or geographical origin. To explore whether such factors can also be traced on the basis of the use of derivational affixes, we narrowed our focus down to the novels by Austen, Bronte, Burroughs, Chu, Conrad, Dickens, Doyle, James, London, Melville, Montgomery, Morris, Orczy, Stoker, Trollope, Twain and Wells. The analysis is based on the same set of affixes.

A principal components analysis reveals five significant dimensions. As before, the first dimension (46.3%) is fully correlated with *-ly* ($r_{\text{-ly},1} = 0.99$). The second dimension (12.7%) appears to be linked with affixes yielding abstract nouns: *-ity* ($r_{\text{-ity},2} = 0.75$), *-ness* ($r_{\text{-ness},2} = 0.55$), and *-ism* ($r_{\text{-ism},2} = 0.62$). However, *-ation* ($r_{\text{-ation},2} = 0.64$) is also correlated with this dimension. The affixes *-ation* ($r_{\text{-ation},3} = -0.68$), and *-ness* ($r_{\text{-ness},3} = 0.67$) appear most strongly on the third dimension (11.2%). Figure 6 shows that the second and third component jointly separate the germanic from the latinate affixes. Even within a single text type, stratal differences between affixes can be traced, although a single dimension no longer suffices. The fourth dimension (6.4%) singles out the use of the superlative suffix ($r_{\text{-est},4} = 0.79$) and to some extent the suffixes *-ful* ($r_{\text{-ful},4} = 0.50$) and *-ize* ($r_{\text{-ize},4} = -0.52$). The fifth component (5.1%) is sensitive to the use of agentive *-er* ($r_{\text{-er},4} = -0.56$) and *-y* ($r_{\text{-est},5} = 0.68$).

As before, we may investigate whether the novels under consideration cluster together in an interpretable way. Although the number of texts is too small to allow any conclusions to be drawn with certainty, some tendencies suggesting a diachronic factor can be observed. Consider figure 7, which plots the texts in the three-dimensional space defined by the principal components one, four and five. Authors born after 1850 have been marked with a hash mark (#). These authors tend to score low on the fourth principal component. They also account for the highest values on the first principal component, and, with the exception of Stoker's *Dracula* (LSd), the same holds for the fifth principal component. Considered jointly, as in figure 8 — the large squares represent the authors born after 1850 — a reasonably consistent pattern emerges, the only glaring exception being Montgomery's *Anne of Avonlea* (LMa#), where country of origin (Canada) and sex[6]

[6] On the dimension where this novel is exceptional (4), Austen's *Pride and Prejudice* also scores high.

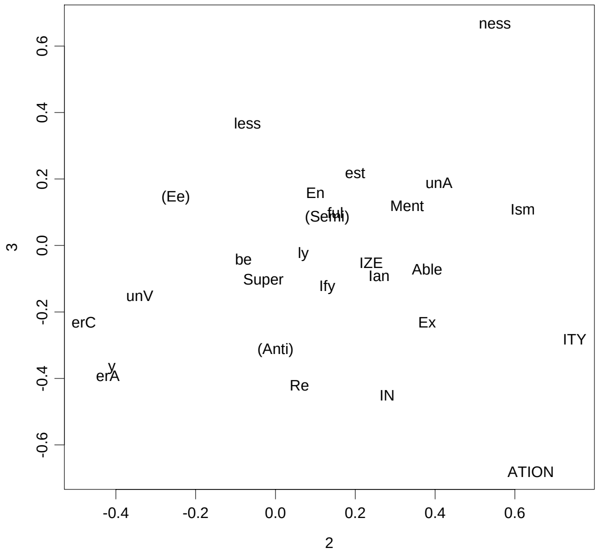

Figure 6: Affixes plotted as a function of their correlations with the second and third principal components in the sample of novels. Affixes that attach to nonnative base words only are shown in upper case, affixes with a latinate origin that attach to both native and nonnative base words have an initial capital letter. The 'germanic' affixes appear in lower case. Affixes with a negligible degree of productivity ($\mathcal{P}^* < 0.05$) are printed between parentheses.

may play a role. Although any conclusions are tentative at best — especially as a number of the authors born after 1850 are represented by more than one text — the major role of the fourth principal component suggests that especially the suffixes *-est* and *-ful* were used more productively by the majority of authors born before 1850, while the suffix *-ize* appears to have been used more productively by the authors born after 1850.

As before, we complement this analysis with a study of the function words. Seven significant principal components were obtained, the first two of which are relevant here. The first component (22.7%) is strongly correlated with *that* ($r_{\text{that},1} = 0.82$), *be* ($r_{\text{be},1} = 0.82$), *but* ($r_{\text{but},1} = 0.80$), *have* (${}_{\text{have},1} = 0.78$), and to some extent *had* ($r_{\text{had},1} = 0.51$). The positive correlations of both *have* and *had* suggest that tense is not singled out by this component. Probably, this component is sensitive to the use of the verb *to have* as such, in combination with the use of subordinate and relative clauses with *that*. The second component (15.7%) is associated with the finite verbal forms of the verb *to be*: *are* ($r_{\text{are},2} = -0.75$), *were* ($r_{\text{were},2} = 0.73$), *is* ($r_{\text{is},2} = -0.72$), and *was* ($r_{\text{was},2} = 0.65$). The positive correlations for past tense forms and the negative correlations for the present tense forms shows that this component registers differences in tense. Other words scoring high on this dimension are *we* ($r_{\text{we},2} = -0.66$) and *as* ($r_{\text{as},2} = 0.61$). Figure 9 plots the texts in the plane defined by these two principal components. Note that authors born before 1850 tend to cluster in the upper right hand corner. There are some exceptions to this pattern, notably the novels by Bronte (LBw) and James (LJe, LJc). Possibly, James' early traveling and subsequent settling in England have caused him to be more sensitive to the English of the last quarter of the nineteenth century than his year of birth (1843) would suggest. Although no definite conclusions can be drawn in the light of the small number of texts figuring in these analyses, the fact that both the morphology-based analysis and the analysis based on function words suggest a development through time shows that more extensive analyses along these lines are potentially rewarding.

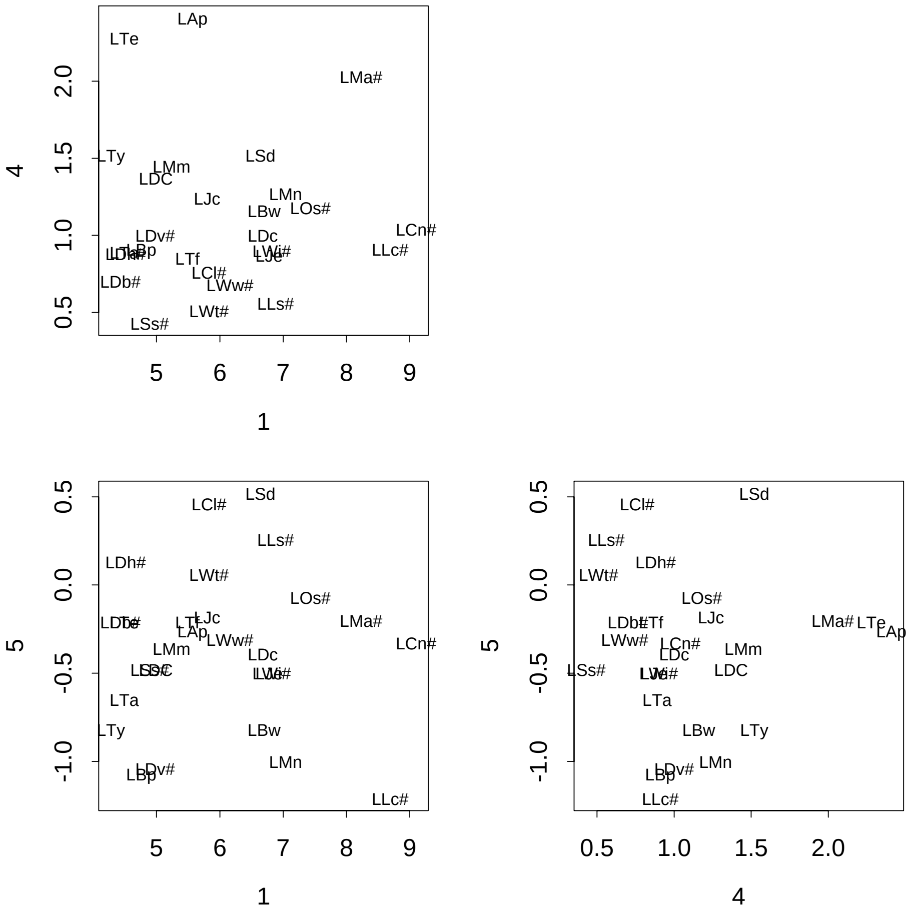

Figure 7: Morphology-based principal components analysis for 27 novels. The scatterplots chart the three-dimensional space spanned by the first, fourth and fifth significant principal components. The three panels can be viewed as the top, the front and the right-hand side of a transparent cube in which the texts are located. The first letter of the codes denote the text type. Texts written by authors born after 1850 are marked with a hash mark (#). For further details see the appendix.

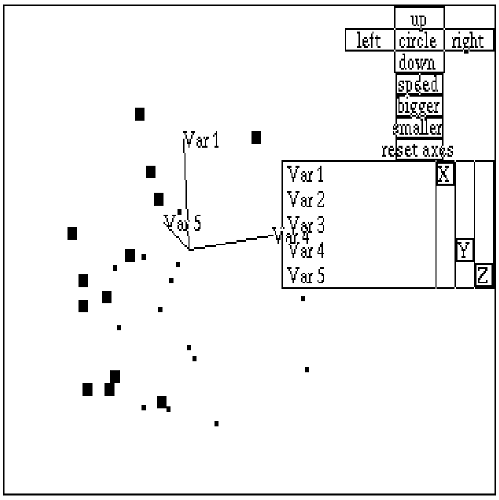

Figure 8: Location of 27 novels in the three-dimensional space defined by the first and last two significant dimensions of a morphology-based principal components analysis. The large squares represent novels written by authors born after 1850.

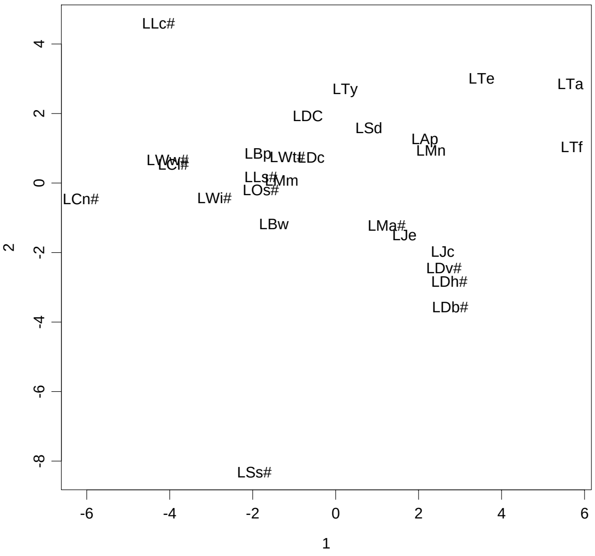

Figure 9: Location of 27 novels in the plane defined by the first two significant dimensions of a function word based principal components analysis. Texts written by authors born after 1850 have been marked with a hash mark (#). For further details see the appendix.

7 Discussion

We have seen that the productivity statistic $\mathcal{P}^*$ allows texts to be grouped together in meaningful clusters. Similar clusterings can be obtained on the basis of the relative frequencies of function words, showing that the observed groupings are robust. In this section we consider the theoretical consequences of this finding for the study of morphological productivity and for literary studies.

For the domain of literary studies, the present results are of interest in that they illustrate that morphology constitutes a fruitful domain of inquiry. Summary plots such as figure 10 can be used to study author-specific preferences for particular affixes. Some texts, such as Orczy's *The Scarlet Pimpernel* (LOs) and London's *The Call of the Wild* (LLc) show no clear preferences for any particular affix. On the other hand, two of Trollope's novels (LTf, LTa) are characterized by an intensive use of *-ness*, Austen (LAp) and James (LJc, LJe) make heavy use of *in-*, *-ity* and *-ation*, and Montgomery (LMa) and Morris (LMn), but also Doyle's

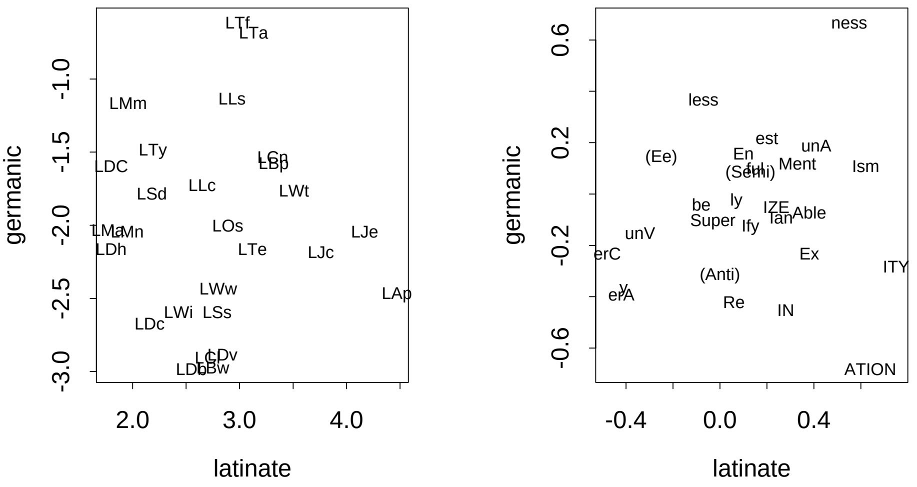

Figure 10: Scatterplots of texts in the plane defined by the latinate and germanic principal components (left), and of affixes in the plane of the corresponding correlation coefficients (right), for the sample of novels. Affixes that are [-native, +learned] are shown in upper case, affixes with the features [+native,+learned] in lower case with an initial capital letter, and [+native, -learned] affixes in lower case. Affixes with a negligible degree of productivity ($\mathcal{P}^* < 0.05$) appear between parentheses.

The Hound of the Baskervilles (LDh) show a marked preference for comparative and agentive *-er* and *-y*. Depending on one's sample of texts, the specific morphological characteristics of authors (in samples of homogeneous texts) or text types (in samples of heterogeneous texts) can be traced.

It should be noted that the morphology-based analyses show that different texts written by a single author do not always cluster together. For instance, Trollope's *The Esutace Diamonds* (LTe) patterns differently from his other two novels, both in the complete and the restricted sample (see figures 7 and 10). Similarly, Doyle's novels (LDh, LDb, LDv) span almost the entire range of the fifth dimension in figure 7. Apparently, there may be substantial fluctuations in the way a single author uses his affixes. Syntactic patterns, as measured through the relative frequencies of function words, appear to be more stable: in the corresponding analysis using function words, texts by a single author tend to cluster more closely (see figure 9). This suggests that for studies of authorship attribution, the function words should be studied as they occur in a sample of stylistically homogeneous texts. For heterogeneous texts, however, the morphology-based approach appears to yield slightly better results (compare figures 3 and 5). Future research will have to clarify whether the present morphology-based analyses are supported by analyses of syntactic, semantic and pragmatic variables along the lines of Biber (1989).

Turning to the domain of linguistics, it is clear that the degree of productivity of a word formation rule is strongly influenced by text type and author-specific preferences. Adverbializing *-ly* is the most productive affix studied here, nevertheless it is not the most productive affix in each and every text. Officialese tends to use *-ly* more sparingly than the majority of novels. And whereas Barrie's *Peter Pan* is characterized by an extremely prolific use of *-ly* ($\mathcal{P}^* = 0.106$), Milton's *Paradise Lost* hardly uses *-ly* at all ($\mathcal{P}^* = 0.017$). In fact, Milton uses adjectivizing *un-*, superlative *-est*, and *-ation* more productively than *-ly*. Similarly, *-ly* is not the most productive suffix in *The Federalist Papers*, where *-ation* is fractionally more productive ($\mathcal{P}^* = 0.046$) than *-ly* ($\mathcal{P}^* = 0.044$). These remarkable inversions in the degree of productivity for *-ly* show that text type and individual preferences are factors that may have more weight than the structural restrictions defining the input domain of a word formation rule.

There are a number of more specific issues requiring some discussion. First consider the distribution of native and nonnative affixes over the principal components. Figure 11 plots affixes in the plane defined by their correlations with the second and third principal components for the complete sample of 44 texts.[7] We may distinguish between three classes of affixes. First, we have affixes that are [+native] in that they attach to both native and nonnative base words. Thus we have *fairness*, *unwise* and *thickly* side by side with *completeness*, *uncertain* and *conspicuously*. At the same time, using Bloomfield's (1933) terminology, these affixes can be characterized as 'non-learned'. In figure 11 (and the figures 6 and 10) these affixes are printed in lower case. Second, there are affixes that, although from latinate origin, attach freely to both latinate and germanic base words (*reforest, reconsider*; *Brownian, Episcopalian*; *workable, retractable*). These affixes, however, are 'learned'. They are also printed in lower case, with the exception of the initial letter. Third, there are learned affixes that attach to latinate base words only (*-ity, -ation, in-*, and *-ize*). They are printed in upper case.

The first thing to note is that all [+learned] affixes have a positive correlation coefficient with the second principal component. Of the four affixes that are [+learned] and [-native], three occur at the right hand edge of figure 11. Conversely, the [-learned, +native] affixes tend to score low on the second principal component, de-adjectival *un-* being the only exception. In addition, the [-learned, +native] affixes show up with both large positive and large negative correlation coefficients for the third principal component. Taken together, we may conclude that the learned affixes, but not the non-learned ones, pattern together as a group. When a particular text type favors the use of learned affixes, all such affixes are used more productively. The affixes that are [-learned, +native] do not cohere in the same way: some are reasonably productive (*-y, -est*) without correlating strongly with any principal component ($|r_i| < 0.40, i = 1, 2, 3$), others are productive, but appear at opposite ends of a single dimension (*-ness, -er*).

[7] Some caution is required for the interpretation of this plot, as the Euclidean distance between two affixes does not always imply a high degree of correlation in their use. Inspection of the Pearson product-moment correlation coefficients and the corresponding t-values shows that all latinate affixes at the right hand side are all significantly correlated ($p < 0.05$). Significant correlations that do not involve the latinate set are discussed separately below.

mean $\mathcal{P}^*$ ($*10^{-2}$)	affix	native	learned	component	$r_{\text{affix, component}}$
0.06	super	+	+	2	0.31
0.08	ian	+	+	1	-0.35
0.08	ism	+	+	2	0.56
0.10	ify	+	+	2	0.20
0.13	be	+	−	2	-0.27
0.14	ize	−	+	2	0.50
0.17	unV	+	−	2	-0.27
0.34	en	+	+	2	0.33
0.36	re	+	+	3	-0.39
0.43	ex	+	+	2	0.51
0.45	able	+	+	2	0.52
0.52	less	+	−	3	0.29
0.61	ity	−	+	2	0.76
0.62	ful	+	−	1	0.23
0.69	ment	+	+	2	0.53
0.84	in	−	+	2	0.78
0.93	est	+	−	2	-0.36
1.05	erC	+	−	3	-0.54
1.11	ness	+	−	3	0.86
1.31	erA	+	−	3	-0.60
1.45	unA	+	−	2	0.66
1.54	y	+	−	1	0.39
2.42	ation	−	+	2	0.90
6.16	ly	+	−	1	1.00

Table 2: Degree of productivity and the maximal correlation coefficients for the affixes with non-negligible degree op productivity ($\mathcal{P}^* > 0.05$) in the full sample of 44 texts.

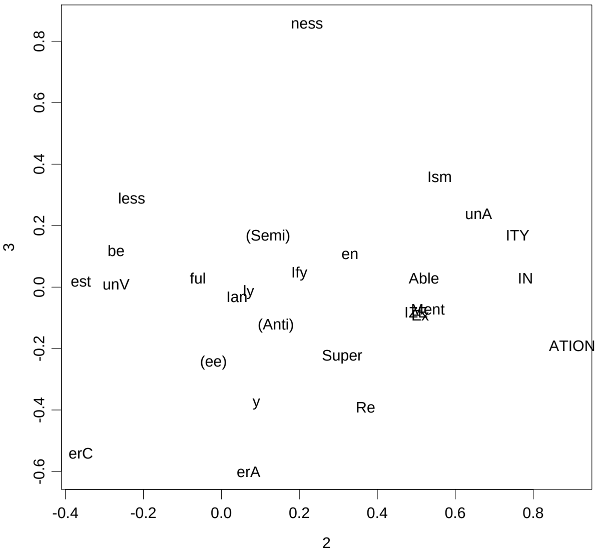

Figure 11: Affixes plotted as a function of their correlations with the second and third principal components. Affixes that are [-native,+learned] are shown in upper case, affixes with the features [+native,+learned] in lower case with an initial capital letter, and [+native, -learned] affixes in lower case. Affixes with a negligible degree of productivity ($\mathcal{P}^* < 0.05$) appear between parentheses.

This difference between the native and nonnative affixes ties in with a difference in markedness. Of the affixes under consideration here, the [+native, -learned] affixes are fully unmarked. The [+native, +learned] affixes are marked, the [-native,+learned] affixes are doubly marked. Figure 11 shows that an increase in markedness goes hand in hand with a higher correlation with the second principal component. The fully unmarked affixes appear at the left hand side, the marked affixes occupy the intermediate range, and the highly marked affixes appear at the extreme right. Recalling that on any level of analysis unmarked elements are more general and have the wider distribution, we may expect unmarked affixes to be less variable across different text types: they are basic to any style. Conversely, marked affixes require a more specialized use. In analyses of a variety of text types, they should evidence the greater variability. With the exception of adverbial -*ly*, this is exactly what we find for our complete sample of 44 texts: the fully unmarked affixes are linked with the third principal component, which explains 7.3% of the variance, while the second principal component, which accounts for 18.1% of the variance, is associated with the marked and highly marked affixes. Within a single text type the use of the marked affixes may be expected to be less variable. This expectation is born out by the analysis of the sample of novels, where two components are now required to separate the marked from the unmarked affixes (see figure 6).

The exceptional behavior of unmarked -*ly*, singly explaining 51.2% of the variance, remains to be accounted for. When we calculate the coefficient of variation for the degree of productivity of -*ly* we find that it is much lower (0.295) than that of -*ness* (0.519). At the same time, -*ly* is very much more productive than all other affixes studied here. This suggests that the variability in the use of -*ly* is magnified out of proportion due to its high degree of productivity, and that it does not constitute a counterexample to the hypothesis that marked affixes explain a larger proportion of the variance.[8] For the other affixes, the degree of productivity is more commensurable. It is not the main determinant of the way affixes group together.

[8] When the analysis is carried out on the covariance matrix instead of on the correlation matrix, similar results are obtained, but now the latinate affixes are represented on the first principal component and the germanic affixes on the second. It is only on the third principal component that -*ly* appears.

Table 2, which lists the affixes with a non-negligible degree of productivity, their average $\mathcal{P}^*$-value, their specifications for the features [native] and [learned], the components with which they are most strongly correlated, and the corresponding correlation coefficients, shows that the average degree of productivity cannot be used to predict the component on which an affix will appear. The suffix *-y* is quite productive, yet it is only weakly correlated with the first principal component of the complete sample of texts. Similarly, the suffixes *-ness* and *-er* are more productive than *-able* and *-ity*. Nevertheless, they correlate with the third component rather than the second.

A second issue requiring some discussion is the patterning of so-called rival (i.e., nearly or fully synonymous) affixes. Our set of affixes includes two such pairs: *-ness* and *-ity*, and *un-* and *in-*. First consider *-ness* and *-ity*. In the complete sample of 44 texts the rival suffixes *-ness* and *-ity*, studied in detail in Aronoff (1976) and analyzed quantitatively in Baayen and Lieber (1991), appear with quite different correlation coefficients for the second and third principal components, as shown in figure 11. At first sight, they pattern rather as unrelated affixes. Nevertheless, inspection of the Pearson product-moment correlation coefficient for the $\mathcal{P}^*$-values for the two affixes reveals a small but non-negligible correlation. ($r = 0.30, p < 0.05$). In fact, there is one other affix with which *-ness* is correlated positively, *-ism* ($r = 0.38, p < 0.02$), an affix that also creates abstract nouns. In addition, *-ness, -ity* and *-ism* evidence a significant negative correlation with comparative *-er*. (The only other affix to do so is *-ify*). This suggests that we are measuring the effects of a semantic factor: once an author is in the mood to coin abstract nouns, the productivity of all three affixes is enhanced. At the same time, we appear to be dealing with three independent suffixes, each with its own semantics (see Riddle (1984) and Romaine (1983) for a number of subtle but important differences in the semantics of *-ness* and *-ity*). Since *-ity* is one of the stratally marked affixes that as such is subject to stylistic forces in quite different ways then *-ness*, we are forced to conclude that *-ness* and *-ity* are rival affixes in a very loose sense only.

Next consider the rival prefixes *un-* and *in-*. The analysis of the complete data set showed *un-* to be the single native affix that is highly correlated with the second principal component. Figure 11 shows that it also has a highly similar value for its correlation coefficient with the third principal component. In contrast to *-ness* and *-ity*, these affixes show up with a fairly high Pearson product-moment correlation coefficient ($r = 0.615$, $p < 0.001$). Interestingly, there is no difference in 'conceptual' meaning between the affixes — both express the negation of some quality or property. Apparently, the use of affixal negation, as opposed to periphrastic negation with *not*, is stylistically marked, *un-* being used more productively in texts favoring the use of learned affixes. Interestingly, *in-* and *un-* move apart in the same way as *-ness* and *-ity* when we factor out gross differences in style, as in the analysis of the set of novels (figure 6). Here *un-* sides with the [+native, -learned] affixes. Once the use of affixal negation as such has become unmarked, the use of [-native, +learned] *in-* surfaces as marked with respect to [+native, - learned] *un-*.[9]

Figure 11 shows that texts making a heavy use of *-ness* tend to use comparative and agentive *-er* sparingly, and vice versa. The Pearson product-moment correlation coefficients are low ($r_{\text{-ness, -er (a)}} = -0.36$, $r_{\text{-ness, er (c)}} = -0.36$, but significant at the 2% level. Along with *-ness*, the latinate suffixes *-ism* and *-ity* also correlate negatively with comparative *-er* ($p < 0.05$). This pattern probably arises due to thematic and semantic constraints on the productivity of affixes: texts for which abstract properties are important are less likely to focus on acting subjects nor, apparently, on the degree to which a quality can be predicated.

Finally, we have seen that age is yet another factor that may co-determine the productivity of a rule. The analysis of the sample of novels suggests that the productivity of the superlative suffix *-est* has decreased over time whereas that of *-ize* increased, the oldest text (Austen's *Pride and Prejudice*) and the youngest text (Chu's *More than a Chance Meeting*) appearing at opposite extremes. As mentioned in section 2, Romaine's (1983) sociolinguistic study of preferences for *-ness* and *-ity* also revealed a subject's age to co-determine her or his productivity judgements. Unfortunately, Romaine's experimental task requires subjects to judge the acceptability of words out of context, where none of the factors that normally shape one's choice of words are present. We have seen that what may be acceptable for one text type may be odd for another kind of text.

[9] There is one example of affixal homonymy in our data, agentive and comparative *-er*. These suffixes have rather similar correlation coefficients for all principal components of the full sample. This, however, appears to be a coincidence. There is no trace of a correlation between their $\mathcal{P}^*$-values ($r = 0.181$, $p > 0.20$).

Hence it is not entirely clear what kind of ability is measured in her experiment. Once enough text materials of spoken and written language become available in electronic form, the present methodology is likely to yield a more profound insight into the sociolinguistic aspects of productivity than in vitro experiments, however useful these may be.

To conclude, it is remarkable to find that analyses based on words from opposite ends of the frequency spectrum — the morphologically complex hapax legomena on the one hand, and the highest frequency (monomorphemic) function words on the other — tend to converge. This suggests that the combined quantitative analysis of morphology on the one hand (in terms of the productivity of various affixes) and syntax and pragmatics on the other (by means of the relative frequency of function words in combination with an analysis of the variables studied by Biber (1989)) constitutes a robust and fruitful line of inquiry into the sociolinguistic and stylistic aspects of language use.

APPENDIX

Author	Title	Source	Birth/Death	Code
Aesop	Fables (translated by G. F. Townsend)	PG	-	CAs
Anonymous	The Federalist Papers	PG	-	OAf
J. Austen	Pride and Prejudice	PG	(1775–1817)	LAp
J. M. Barrie	Peter Pan and Wendy	OTA	(1860–1937)	CBp
L. F. Baum	The Wonderful Wizard of Oz	PG	(1856–1919)	CBw
L. F. Baum	The Marvelous Land of Oz	PG	(1856–1919)	CBo
E. Bronte	Wuthering Heights	PG	(1818–1848)	LBw
E. R. Burroughs	A Princess of Mars	OTA	(1837–1921)	LBp
L. Carroll	Alice's Adventures in Wonderland	PG	(1832–1898)	CCa
L. Carroll	Through the Looking Glass and what Alice Found there	PG	(1832–1898)	CCt
B. Clinton	Election Speeches	OBI (?)	(1945 -) (?)	OCl
J. Conrad	Lord Jim	OTA	(1857–1924)	LCl
J. Conrad	Nigger of the Narcissus	OTA	(1857–1924)	LCn
C. Darwin	On the Origin of the Species	OTA	(1809–1882)	ODo
C. Dickens	A Christmas Carol	PG	(1812-1870)	LDC
C. Dickens	The Chimes: a Goblin Story	OTA	(1812–1870)	LDc
A. C. Doyle	The Hound of the Baskervilles	PG	(1859–1930)	LDh
A. C. Doyle	The Casebook of Sherlock Holmes	OTA	(1859–1930)	LDb
A. C. Doyle	The Valley of Fear	OTA	(1859–1930)	LDv
Government Accounting Office	Selected Texts	OBI		OGa
J. & W. Grimm	Fairy Tales (Translations)	OBI		CGr
Congress Hearings	Selected Texts	PG		OCh
H. James	Confidence	OTA	(1843–1916)	LJc
H. James	The Europeans	OTA	(1843–1916)	LJe
W. James	Essays in Radical Empiricism	OBI	(1842–1910)	OJe
R. Kipling	The Jungle Book	PG	(1865–1936)	CKj
J. London	The Sea Wolf	OTA	(1876–1916)	LLs
J. London	The Call of the Wild	OTA	(1876–1916)	LLc
King James Version	Luke-Acts	PG		BLu
H. Melville[†]	Moby Dick	PG	(1819–1891)	LMm
J. Milton	Paradise Lost	PG	(1608–1674)	LMp
L. M. Montgomery	Anne of Avonlea	OTA	(1874–1942)	LMa
J. Smith	The Book of Mormon	OBI	(1805–1844)	BMo
W. Morris	News from Nowhere	OTA	(1834–1896)	LMn
E. Orczy	The Scarlet Pimpernel	OTA	(1865–1947)	LOs
C. C. Chu	More than a Chance Meeting (Startrek) (Startrek)	OBI	-	LSs
B. Stoker	Dracula	OTA	(1847–1912)	LSd
A. Trollope	The Eustace Diamonds	OTA	(1815–1882)	LTe
A. Trollope	Can you Forgive her?	OTA	(1815–1882)	LTf
A. Trollope	Ayala's Angel	OTA	(1815–1882)	LTa
M. Twain	A Connecticut Yankee in King Arthur's Court	OTA	(1835–1910)	LTy
H. G. Wells	The Time Machine	PG	(1866–1946)	LWt
H. G. Wells	The War of the Worlds	PG	(1866–1946)	LWw
H. G. Wells	The Invisible Man	OTA	(1866–1946)	LWi

[†] The electronic text of *Moby Dick* was prepared by E.F.Tray at the University of Colorado, Boulder, on the basis of the Hendricks House edition.

References

Antworth, E.L. (1990), PC-KIMMO*: a two-level processor for morphological analysis.* Dallas: Summer Institute of Linguistics.

Aronoff, M. (1976), *Word Formation in Generative Grammar.* Cambridge, Mass.: The MIT Press.

Baayen, R.H. (1989), *A Corpus-Based Approach to Morphological Productivity. Statistical Analysis and Psycholinguistic Interpretation.* Diss. Vrije Universiteit, Amsterdam.

Baayen, R.H. (1992), "A Quantitative Approach to Morphological Productivity". In *Yearbook of Morphology 1991*, Eds. G.E.Booij and J.van Marle. Dordrecht: Kluwer, 109–149.

Baayen, R.H. (1993), "On frequency, transparency and productivity". In *Yearbook of Morphology 1992*, Eds. G.E.Booij and J.van Marle. Dordrecht: Kluwer, 181–208.

Baayen,R.H. (1993b), "Statistical Models for Word Frequency Distributions: A Linguistic Evaluation". *Computers and the Humanities* 26: 347–363.

Baayen, R.H. & Lieber, R. (1991), "Productivity and English Derivation: a corpus-based study". *Linguistics* 29:801–844.

Becker, R.A., Chambers, J.M. and Wilks, A.R. (1988), *The New S Language. A programming environment for data analysis and statistics.* Pacific Grove: Wadsworth & Brooks/Cole.

Biber, D. (1989), "A Typology of English Texts". *Linguistics* 27, 3–43.

Biber, D. and Finegan, E. (1989), Drift and Evolution of English Style: A History of Three Genres. *Language* 65, 487–414.

Bloomfield, L. (1933), *Language.* London: Allen and Unwin.

Booij, G.E. (1977), *Dutch Morphology. A Study of Word Formation in Generative Grammar.* Dordrecht: Foris.

Burgschmidt, E. (1977), "Strukturierung, Norm und Produktivität in der Wortbildung". In *Perspektiven der Wortbildungsforschung.* Eds. H.E.Brekle and D. Kastovsky. Bonn: Bouvier Verlag, 39–47.

Burrows, J.F. (1992), "Computers and the Study of Literature". In *Computers and Written Texts.* Ed. C.S.Butler. Oxford: Blackwell, 167–204.

Burrows, J.F. (1993), "Noisy Signals? or Signals in the Noise?" In *ACH-ALLC Conference Abstracts*, Georgetown, 21–23.

Chitashvili, R.J. & Baayen, R.H. (in press), "Word Frequency Distributions". In: L. Hřebíček & G.Altmann (Eds.), *Quantitative Text Analysis.* Trier: Wissenschaftlicher Verlag.

Haeringen, C.B. van (1971), "Het achtervoegsel -ing. Mogelijkheden en beperkingen". [The suffix -ing. Possibilities and Restrictions.] *De Nieuwe Taalgids*, 64, 449–468.

Morrison, D.F. (1976), *Multivariate Statistical Methods.* Tokyo: McGraw-Hill Kogakusha.

Rainer, F. (1988), "Towards a Theory of Blocking: the Case of Italian and German Quality Nouns". In *Yearbook of Morphology 1*, Eds. G.E.Booij and J.van Marle. Dordrecht: Foris, 155-185.

Riddle, E.A. (1984), "A Historical Perspective on the Productivity of the Suffixes -ness and -ity". Paper presented at the *Conference on Historical Semantics and Word Formation, Blazejenko, Poland*, 28–31.

Romaine, S. (1983), "On the Productivity of Word Formation Rules and Limits of Variability in the Lexicon". *Australian Journal of Linguistics* 3, 177–200.

Santen, A. van (1992), *Produktiviteit in taal en taalgebruik.* Leiden.

StatSci (1991), *S-Plus User's Manual, Vol.1, Version 3.0.* Headington: Statistical Sciences U.K. Ltd.