# Socio-geographic variation in morphological productivity in spoken Dutch : a comparison of statistical techniques.

Karen Keune[1], Roeland van Hout[1], R. Harald Baayen [1,2]

1Radboud University of Nijmegen, the Netherlands

[2]Max Planck Institute for Psycholinguistics, the Netherlands

## Abstract

This study explores socio-geographic variation in morphological productivity in spoken Dutch. For 72 affixes, we extracted the hapax legomena from the Corpus of Spoken Dutch. We divided the corpus into 24 subcorpora defined by the speaker's country (Flanders versus The Netherlands), education level (High versus Non-High), sex (Women versus Men), and age (Young, Mid or Old). The large number of cells with zero counts for the affixes, and the substantial variation in the sizes of the subcorpora underlying the cell counts, posed a special challenge for the statistical analysis. We fitted three different kinds of models to our data : an ordinary least squares linear model with a transformation of the proportion of hapax legomena in the subcorpus as dependent variable, a linear mixed effects model with affix as random effect and the transformed proportions as dependent variable, and a generalized linear model with a binomial link which considered the hapax legomena as successes, and all remaining words as failures. The generalized linear model outperformed the others, in spite of the extremely small probabilities of success. We discuss why the generalized linear model is superior, and show how generalized linear models can be used to visualize by-affix variability in productivity.

**Keywords** : socio-geographic variation, generalized linear models, morphology, productivity, visualization.

## 1. Introduction

According to Bauer (2001), an affix is productive if it is possible to create new words with it. An example of a productive affix in English is *-ness*. One can easily form new words ending in *-ness*. For instance, given the adjective *fractionated*, one can form the noun fractionatedness. By contrast, the suffix *-th*, as in *warmth*, is hardly productive, although an occasional neologism can be observed (Baayen, 2003).

There are several quantitative measures available for gauging the degree to which an affix is productive. An obvious measure is the size of the set of words containing the affix, henceforth its morphological category, as observed in a corpus. The more words an affix attaches to, the more productive that affix is. The disadvantage of this measure is that it does not take into account possible diachronic change in the productivity of an affix. So, a morphological category like that of the suffix *-ment*, which was more productive in the past, still has a considerable number of members, even though modern speakers are reluctant to use it in new words (Anshen and Aronoff, 1999). Conversely, speakers may also be reluctant to use an affix, even though it is fully productive in the sense that they could use it if required. An example is the Dutch suffix *-ster*, used to create nouns referring to female agents, such as *loop-ster*, 'female walker'. It is not fashionable in current Dutch to make the sex of the agent explicit, and the use of the unmarked counterpart with the suffix *-er* is preferred instead (Baayen, 1994b).

In order to overcome these difficulties, measures based on the Good-Turing estimate for unseen species (Good, 1953) have been introduced (Baayen, 1993). The measures that we will use here, first proposed in Baayen (1993), estimate the likelihood of observing a new formation with a given affix by counting the number of words that are observed only once, the hapax legomena, and calculating the proportion of such words with that affix. Since hapax legomena are relatively often new words and the number of new words created with a certain affix determines the productivity of that affix, hapax legomena are suited to predict the current productivity of affixes. Instead of measuring the extent to which a morphological category has been used in the past, this measure estimates the rate at which a morphological category is expanding and attracting new members.

Baayen and Renouf (1996) showed, on the basis of a large corpus study of British English, that hapax legomena are indeed the best estimators for the use of neologisms. Nishimoto (2003) compared productivity rankings obtained with the Good-Turing estimate with productivity rankings based on the deleted estimation method of Jelinek and Mercer (1985), and obtained similar rankings for both measures.

Most studies on productivity have proceeded on the implicit assumption that there would be an ideal speaker in a homogeneous speech community, whose knowledge is representative for all the other speakers in that community. In the study of Bauer (2001), for instance, the possibility of variation in degrees of productivity across registers, social groups, and regions is not considered. Given the fact that such variation is involved in many linguistic variables (Biber, 1995 ; Keune et al., 2005), it is likewise expected to be involved in morphological productivity.

Previous variational studies of morphological productivity focused on how productivity varied with text type (Baayen, 1994a) and with register (Plag et al., 1999). Baayen (1994a) found that in some texts, like stories for children, the use of Germanic affixes is preferred, while in more official registers, Latinate affixes are most productive. Plag et al. (1999) showed that the productivity of a suffix may differ between written, formal spoken, and informal spoken language. Suffixes tended to be most productive in written language, and least productive in informal spoken language.

The aim of the present paper is to obtain further insight into the socio-geographic forces shaping morphological productivity in spoken Dutch. We investigated productivity as a function of whether a speaker lives in the Netherlands or in Flanders, of the speaker's sex, education level, and age.

## 2. Materials

We based our study on the Corpus of Spoken Dutch (CGN) (Oostdijk, 2002). This corpus consists of approximately 8.9 million words of spoken Dutch from various speech registers. These can be divided into three main registers, namely, spontaneous speech (unscripted conversations and telephone dialogues, 4.7 million words), speech from more formal settings such as debates, meetings, and interviews (3.3 million words), and read aloud speech of written Dutch (0.9 million words). As we were interested in exploring variation in spoken Dutch, we did not take into account the read aloud speech in the corpus. This left us with a corpus consisting of approximately 8.0 million words.

In the CGN, the characteristics of the speakers, for instance their home country, education level, sex, and age, are made available. This made it possible to address the socio-geographic variation in morphological productivity. To this end, we extracted 24 subcorpora according to a 2x2x2x3 factorial design with as predictors Country (The Netherlands versus Flanders, Education (High versus Non-High), Sex (Men versus Women), and Age (Young : <40 ; Mid : 41-60 ; Old : >60). The size of these subcorpora differed substantially, ranging from 27418 words (for old Flemish male speakers with a non high education level), to 942990 words (for middle aged Dutch male speakers with a high education level).

There are two slightly different criteria for what counts as a hapax legomena with a given affix. One criterion is to include only those hapax legomena with a given affix, for which that affix was attached to the word during the last morphological cycle. According to this criterion, the word *dank-baar-heid* ('grate-ful-ness') would be included in the count for *-heid* but not in the count for *-baar*. Another criterion is to include any word with that affix, including embedded words, as long as these embedded words are not present independently in the corpus either by itself or in other words. According to this criterion, *dank-baar-heid* would be included for the count of *-baar* if and only if *dank-baar* is not observed by itself. Gaeta and Ricca (to appear) have shown that both criteria lead to very similar productivity rankings.

In order to facilitate extraction of the hapax legomena from the (only partially morphologically parsed) corpus, we selected those words in which affixes occurred either in the beginning or at the end of the word. We relaxed the first criterion by allowing words into our counts for which the affix is not attached during the last cycle, but only when they satisfied the second criterion. For instance, *pianospeler* ('piano player') fits our selection criteria given that *speler* is not present independently or as part of another complex word. The inclusion of these words did not substantially influence our results, but helped alleviate the problem of data sparseness.

The different affixes were selected on the basis of their existence in the morphologically parsed part of the CELEX lexical database (Baayen et al., 1993). Existing affixes were only used if there were ten of more word types in CELEX formed with that affix. Next, we restricted ourselves to the use of only those remaining affixes that appeared in the ANS grammar of Dutch (Geerts et al., 1984). In this way, 91 different affixes were selected for further analysis.

In order to determine the number of hapax legomena of these affixes as used in Dutch speech, we selected every word ending in the same characters as the affix from the CGN. So, for instance, for the prefix *be-* we selected al words starting in *be*. We designed a program that decided whether a word was a hapax legomena or was used more frequently, for instance in another inflectional form or as a part of another (often morphological complex) word. This program used the Memory-Based Morphological Analysis parser (Van den Bosch and Daelemans, 1999) that parses morphological complex words. The output of our program consisted of possible hapax legomena. We manually determined whether these words indeed contained the desired affix. For only 72 of the 91 selected affixes, hapax legomena occurred in the data set. In total 2251 hapax legomena were observed. The different affixes and their total number of hapax legomena are displayed in Table 1. In order to be able to measure the productivity of the affixes among the different subcorpora, we determined for each of the 2251 hapax legomena to which of the 24 subcorpora it belonged.

| Affix | Frequency | Affix | Frequency | Affix | Frequency |
|---|---|---|---|---|---|
| aarts- | 1 | -erwijs | 5 | -ie | 24 |
| -ateur | 1 | -ief | 5 | -ist | 24 |
| hyper- | 1 | -es | 6 | -loos | 24 |
| -lijks | 1 | in- | 6 | ont- | 25 |
| opper- | 1 | -nis | 6 | -iteit | 26 |
| pseudo- | 1 | oer- | 6 | -aar | 27 |
| -uur | 1 | -zaam | 6 | -iseer | 27 |
| -dom | 2 | -erd | 7 | -atie | 30 |
| -ent | 2 | -matig | 7 | -erig | 32 |
| -erik | 2 | -air | 9 | -lijk | 34 |
| -st | 2 | -te | 9 | her- | 37 |
| tele- | 2 | -sel | 10 | -ij | 41 |
| -waarts | 2 | -ant | 11 | -baar | 43 |
| -aard | 3 | de- | 11 | be- | 47 |
| -elaar | 3 | -schap | 12 | super- | 61 |
| oud- | 3 | -aal | 13 | on- | 64 |
| psycho- | 3 | -eel | 13 | -isch | 72 |
| -weg | 3 | -ator | 14 | -achtig | 90 |
| -abel | 4 | inter- | 14 | -heid | 100 |
| bio- | 4 | -ling | 14 | ver- | 114 |
| -gewijs | 4 | -ster | 20 | -ing | 141 |
| -in | 4 | -isme | 21 | -er | 175 |
| co- | 5 | ge- | 23 | -ke | 184 |
| dis- | 5 | anti- | 24 | -je | 477 |

*Table 1 : The 72 different affixes and their number of hapax legomena in The Corpus of Spoken Dutch.*

## 3. Method

The collected data posed a special challenge for statistical analysis for several reasons. First, many affixes emerged with zero counts for a large number of cells in the design. Second, each of the cells in the design contained counts based on subcorpora that differed in size by an order of magnitude. In order to illustrate the diversity of the subcorpora, we displayed their size, the hapax frequency of their most productive affix, and the mean and the median of their total number hapax legomena in Table 2.

These size differences are due to the problems encountered by the builders of the Corpus of Spoken Dutch to obtain sufficient materials from non-highly educated speakers. Hence, any analysis based on the counts themselves, without taking the size of the subcorpora into account, would largely reflect the inequalities in the sizes of these subcorpora. Third, we needed to address the question whether to treat Affix as a fixed effect or a random effect. Since our sample is not exhaustive, one might argue that Affix is a random effect. On the other hand, we have sampled the most productive affixes, hence the sample is far from random, and might just as well be treated as fixed.

| Subcorpus | Corpus Size | Max | Mean | Median |
|---|---|---|---|---|
| Nl male H Y | 594692 | 47 | 2.4 | 1 |
| Fl male H Y | 450170 | 25 | 2.1 | 1 |
| Nl male NH Y | 234052 | 12 | 0.7 | 0 |
| Fl male NH Y | 122048 | 10 | 0.4 | 0 |
| Nl female H Y | 831388 | 71 | 2.8 | 0 |
| Fl female H Y | 554560 | 36 | 1.8 | 0 |
| Nl female NH Y | 318888 | 33 | 1.0 | 0 |
| Fl female NH Y | 128470 | 13 | 0.4 | 0 |
| Nl male H M | 942990 | 59 | 4.1 | 1 |
| Fl male H M | 574673 | 24 | 3.7 | 1 |
| Nl male NH M | 178167 | 11 | 0.6 | 0 |
| Fl male NH M | 52833 | 4 | 0.3 | 0 |
| Nl female H M | 481097 | 37 | 1.9 | 0 |
| Fl female H M | 424558 | 20 | 1.8 | 1 |
| Nl female NH M | 169749 | 14 | 0.5 | 0 |
| Fl female NH M | 51483 | 5 | 0.3 | 0 |
| Nl male H O | 344009 | 23 | 2.2 | 1 |
| Fl male H O | 283929 | 21 | 1.7 | 0 |
| Nl male NH O | 93095 | 3 | 0.2 | 0 |
| Fl male NH O | 27418 | 1 | 0.1 | 0 |
| Nl female H O | 166320 | 17 | 0.8 | 0 |
| Fl female H O | 132367 | 16 | 0.7 | 0 |
| Nl female NH O | 182288 | 22 | 0.6 | 0 |
| Fl female NH O | 38865 | 4 | 0.2 | 0 |
| Total | 7378109 | 71 | 1.3 | 0 |

*Table 2 : The size of each subcorpus, the number of hapaxes of the most productive affix in the subcorpus, the mean and the median of the occurrences of the total number of hapax legomena in the subcorpus. Fl = Flanders, Nl = Dutch, H = High educated, NH = Non High educated, Y = aged < 41, M = aged 41-60, O = aged > 60.*

In the light of these challenges, we analyzed the data with three different statistical techniques. Our first model was obtained using ordinary least squares regression with the proportion of hapax legomena in the subcorpus as dependent variable. The statistic formula is given below :

$E[Y] = X\beta + \varepsilon$, in which Y represents the criterion, X the predictors and β the weights.

We rescaled these proportions by multiplying them by 100000, and raised them to the power of 0.25 in order to reduce the skew in their distribution. Since proportions for large subcorpora are more reliable than proportions for small subcorpora, we fitted the ordinary least squares model to the data using the sizes of the subcorpora as weights. In this model, we treated Affix as a fixed effect. The results for a model containing only simple effects are shown in the left section of Table 3, the results for a model in which two-way interactions were allowed are listed in Table 4.

We also analyzed the data with a linear mixed effects model with Affix as random effect, using the same transformed proportions as in the ordinary least squares regression. The statistic formula looks as follows :

$E[Y] = X\beta + Zb + \varepsilon$, in which Y represents the criterion, X the data matrix, β the coefficients of the fixed effects, Z a copy of the data matrix, and b the coefficients of the random effects.

We used the *lme4* library of Bates and Sarkar (2005), using restricted maximum likelihood estimation. The *lme4* library provides improved algorithms compared to the *nlme* library of Pinheiro and Bates (2000), but has the disadvantage that it is still under development. At the time of writing, it was not possible for us to make use of weighted models. The results obtained with this multilevel model are listed in the central sections of Tables 3 and 4.

Our third model made use of a generalized linear model with a binomial link function, of which the statistic formula is :

$E[Y] = 1/ (1+e^{-X\beta})$ in which Y represents the criterion, X the predictors and β the weights.

Hapax legomena were considered as successes, and all remaining words in the subcorpus were counted as failures. In this logistic model, the total number of words in the subcorpora is automatically included as weight. The third sections of Tables 3 and 4 summarize the results obtained. As we were coping with already fairly sparse data, we did not take three-way interactions into account in any of these analyses.

| | Lm | | | lmer | | | glm | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | F | $df_1$ | P | F | $Df_1$ | p | F | $df_1$ | $df_2$ | p |
| Country | 8.07 | 1 | 0.0046 | 0.02 | 1 | 0.8898 | 18.94 | 1 | 1726 | <0.0001 |
| Education | 182.29 | 1 | <0.0001 | 207.02 | 1 | <0.0001 | 22.10 | 1 | 1725 | <0.0001 |
| Sex | 58.93 | 1 | <0.0001 | 21.57 | 1 | <0.0001 | 34.33 | 1 | 1724 | <0.0001 |
| Age | 13.70 | 2 | <0.0001 | 12.04 | 2 | <0.0001 | 16.79 | 2 | 1722 | <0.0001 |
| Affix | 27.87 | 71 | <0.0001 | | | | 64.02 | 72 | 1651 | <0.0001 |
| | $R^2 = 0.42$ | | | $R^2 = 0.45$ | | | $R^2 = 0.73$ | | | |

*Table 3 : F and p statistics for three simple main effects models : an ordinary least squares model (lm), a multi level model (lmer), and a generalized linear model (glm). For lm, $df_2$ = 1651, for lmer, df2 = 1722.*

## 4. Results

All three models revealed highly significant simple main effects for Education Level, Sex, Age, and Affix. In the ordinary least squares model and in the logistic regression the effect for Country was also significant. Highly educated older men revealed the greatest overall productivity. As expected, productivity varied substantially from affix to affix. The prefix *pseudo*- turned out to be least productive, and the diminutive suffix *-je* to be most productive. For each of the three models, we calculated the squared correlation of the observed and expected cell counts. The resulting $R^2$ was largest for the logistic regression model (0.73), and substantially smaller for the other two models (0.42 and 0.45).

When we allowed two-way interactions into the ordinary least squares model (see Table 4), many interactions emerged as significant, and the $R^2$ increased from 0.42 to 0.73. For the multilevel model, the addition of two-way interactions led to only a small improvement in the $R^2$ from 0.45 to 0.49. The generalized linear model with two-way interactions emerged as most successful, with an increase in the $R^2$ from 0.73 to 0.95.

| | Lm | | | lmer | | | glm | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | F | $df_1$ | p | F | $df_1$ | p | F | $df_1$ | $df_2$ | p |
| Country | 10.53 | 1 | 0.0012 | | | n.s. | 18.94 | 1 | 1726 | <0.0001 |
| Education | 237.78 | 1 | <0.0001 | 118.28 | 1 | <0.0001 | 22.10 | 1 | 1725 | <0.0001 |
| Sex | 76.86 | 1 | <0.0001 | 25.22 | 2 | <0.0001 | 34.33 | 1 | 1724 | <0.0001 |
| Age | 17.88 | 2 | <0.0001 | 10.54 | 1 | <0.0001 | 16.79 | 2 | 1722 | <0.0001 |
| Affix | 36.35 | 71 | <0.0001 | | | n.s. | 64.02 | 71 | 1651 | <0.0001 |
| Country :Sex | 4.65 | 1 | 0.0313 | | | n.s. | 5.69 | 1 | 1293 | 0.0170 |
| Country :Age | 9.75 | 2 | <0.0001 | | | n.s. | 4.95 | 2 | 1294 | 0.0070 |
| Educ :Sex | 16.70 | 1 | <0.0001 | 28.27 | 1 | <0.0001 | | | | n.s. |
| Sex :Age | 4.11 | 71 | 0.0167 | | | n.s. | | | | n.s. |
| | F | $df_1$ | p | $\chi^2$ | $df_1$ | p | F | $df_1$ | $df_2$ | p |
| Cntry :Affix | 3.54 | 71 | <0.0001 | 38.42 | 2 | <0.0001 | 5.37 | 71 | 1580 | <0.0001 |
| Educ :Affix | 2.01 | 71 | <0.0001 | 26.97 | 3 | <0.0001 | 2.10 | 71 | 1509 | <0.0001 |
| Sex :Affix | 2.32 | 71 | <0.0001 | | | n.s. | 2.43 | 71 | 1438 | <0.0001 |
| Age :Affix | 1.80 | 142 | <0.0001 | 15.65 | 4 | 0.0079 | 1.80 | 142 | 1296 | <0.0001 |
| | $R^2 = 0.73$ | | | $R^2 = 0.49$ | | | $R^2 = 0.95$ | | | |

*Table 4 : F and p statistics for three models allowing two-way interactions : an ordinary least squares model (lm), a multi level model (lmer), and a generalized linear model (glm). For lm, $df_2 = 1290$, for lmer, $df_2 = 1722$.*

The substantially better fit achieved with the generalized linear model is due to two factors. First, inspection of the residuals shows that the generalized linear model is more successful in predicting the zero counts. The generalized linear model is not constrained by the normality assumption that governs the distribution of the residuals in ordinary least squares regression. Second, the disappointing performance of the linear mixed effect model is due to the Zipfian nature of affix productivity. Linear mixed effect models assume that random effects follow a normal distribution with mean zero and unknown variance. When we include Affix as a random effect in the multilevel model, we implicitly assume that the difference in productivity of a given affix compared to the average productivity of an affix is normally distributed. This distribution, however, is decidedly non-normal. This explains the disappointing performance of the linear mixed effect model : it is simply not appropriate for our kind of data. For the discussion of the interactions, we therefore restrict ourselves to the logistic regression model.

The interaction of Country by Sex ($F (1, 1293) = 5.69$, $p <0.0170$) indicates that in both the Netherlands and Flanders women use affixes less productively then men. The interaction of Country by Age ($F (1,572) = 10.31$, $p <0.0013$) is illustrated in the upper left panel of Figure 1. Affixes were used less productively by speakers aged between 19 and 40 than by speakers above 40 ($F (1, 1723) = 28.52$, $p <0.0001$). The interaction of the subset of speakers with age above 40 was also significant ($F (1, 1790) = 6.89$, $p <0.0087$). While in the Netherlands speakers above 60 use affixes more productively, in Flanders they use them less productively compared to middle aged speakers. In other words, in the Netherlands productivity increases with age, while in Flanders, the old age group is intermediate between the young and middle age group. The relatively low productivity for older speakers in Flanders may be due to the fact that Dutch was not the official language in Flanders until 1963 (Geeraerts et al., 1999). For these speakers, Dutch is somewhat more like an official register in which they are less fluent, and less productive. However old speakers from Flanders use affixes less productively then middle-aged speakers, they still use them more productively then Dutch speakers. This is probably due to the fact that Flemish speakers have an additional vocabulary (Southern Dutch, Flemish), while Dutch speakers only use the standard (Northern) vocabulary (e.g., Geeraerts et al., 1999).

The productivity of the affixes also varied from affix to affix for all four predictors, as witnessed by the interactions of Country by Affix ($F (71,1580) = 5.37$, $p <0.0001$), Education by Affix ($F (71,1509) = 2.10$, $p <0.0001$), Sex by Affix ($F (71,1438) = 2.43$, $p <0.0001$), and Age by Affix ($F (142,1296)=1.80$, $p <0.0001$).

We visualized the interaction of Country by Affix in the upper left and lower left and right panels of Figure 1. Thanks to the use of contrast coding, with contrasts being made between a given affix and the least productive affix (which was *pseudo*-), the coefficients of Affix and of Affix by Country provide a straightforward estimate of differences in degrees of productivity within and across two countries. The plots are calibrated for young highly educated women.

In the upper right panel the two most productive affixes, the diminutives *-je* and *-ke* are clearly differentiated : *-ke* appears in the upper left corner, which means that it is more productive in Flanders, while *-je* appears in the upper right corner, indicating that it is used more productively in the Netherlands. This is exactly as expected, as these two forms of the diminutive are well-known regional markers (Geerts et al., 1984).

The diminutives enjoy the greatest productivity in spoken Dutch of all our affixes. In order to visualize the structure of the cluster in the lower left hand corner, we zoomed in on this part of the plot, resulting in the lower left panel. This panel reveals that the suffixes *-erig*, *-er*, and –

*achtig*, and the prefix *super-* are more productive in the Netherlands, while the prefixes *her-*, *anti-*, *be-*, and *on-* are more productive in Flanders. The lower right panel zooms in on the cluster of least productive affixes. The suffix *-atie* was reported by Pauwels (1964) to be more productive in Flanders, and his conclusion is supported by our data : *-atie* is located above the Y = X line.

In summary, our multivariate approach to variation in morphological productivity succeeds not only in capturing regional differences already known from the previous literature to exist (*-je* versus *-ke*, *-ing* versus *-atie*), but also offers the possibility to explore many potential carriers of socio-geographic variation simultaneously.

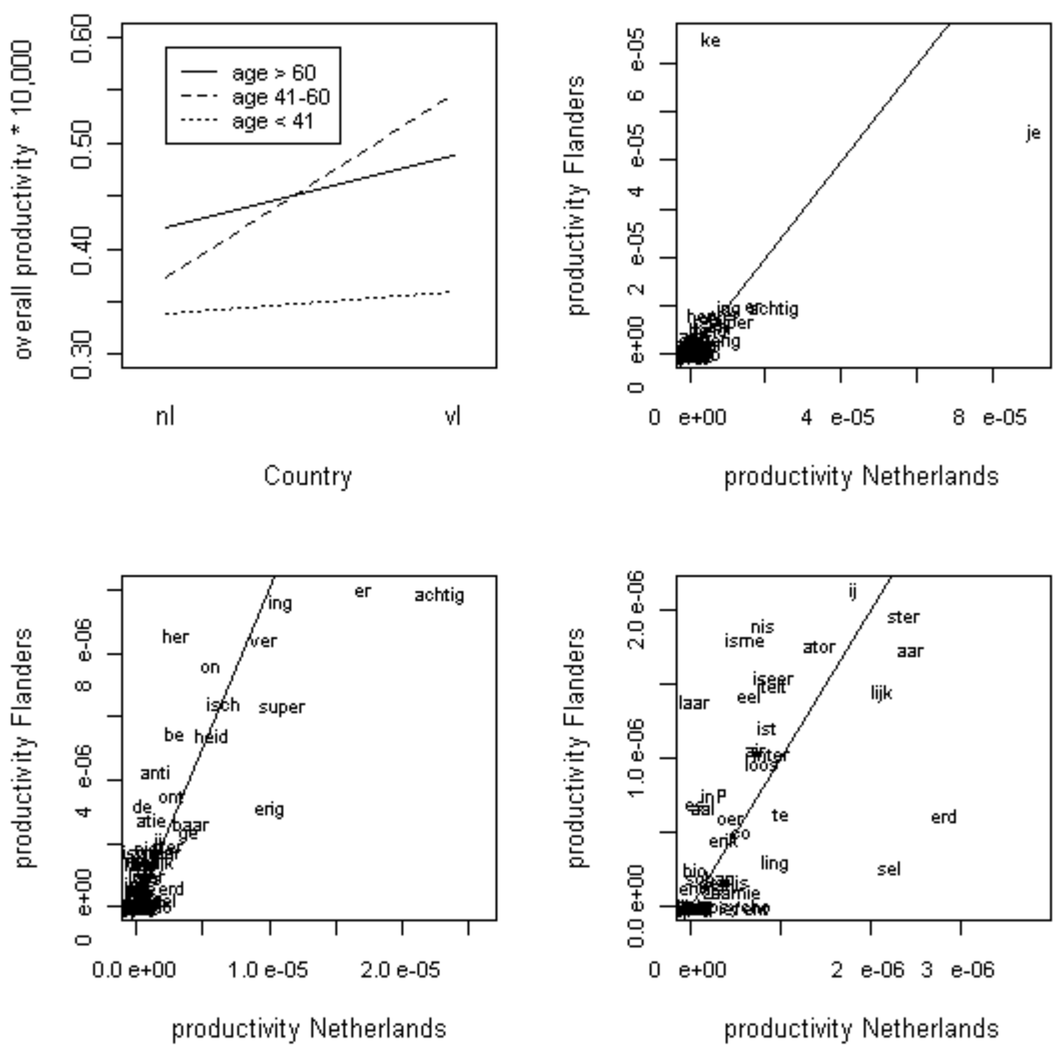


*Figure 1 : The upper left panel illustrates the interaction of country by age. The y-axis depicts the overall productivity of the affixes times 100000. The remaining panels visualize the by-affix adjustments for Country. The x-axes represent the productivity of the affixes for young women from the Netherlands, the y-axes their productivity for young women from Flanders. The lower panels are close ups of the upper right panel.*

## 5. Conclusions

We have shown that it is possible to chart variation in morphological productivity across socio-geographic dimensions, even when there are substantial differences in the sample sizes underlying the counts in the cells of the statistical design. We obtained excellent results with a generalized linear model with a binomial link, even though the success probabilities in our data were extremely small. Given the possibilities for visualization of the variation in the use of the individual affixes, we believe the present approach offers a useful alternative to correspondence analysis for count data in cells with different underlying sample sizes.

## References

Anshen, F. and Aronoff, M. (1999). Using dictionaries to study the mental lexicon. *Brain and Language*, vol.(68) : 16–26.

Baayen, R. H. (2003). Probabilistic approaches to morphology. In Bod, R., Hay, J. and Jannedy, S., editors, *Probabilistic linguistics*, The MIT Press : 229–287.

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

Baayen, R. H. (1994a). Derivational productivity and text typology. *Journal of Quantitative Linguistics*, vol.(1) : 16–34.

Baayen, R. H. (1994b). Productivity in language production. *Language and Cognitive Processes*, vol.(9) : 447–469.

Baayen, R. H., Piepenbrock, R. and Van Rijn, H. (1993). *The CELEX lexical database* (CD-ROM). Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Baayen, R. H. and Renouf, A. (1996). Chronicling The Times : Productive Lexical Innovations in an English Newspaper. *Language*, vol.(72) : 69–96.

Bates, D. M. and Sarkar, D. (2005). *The lme4 library*. [On-line], Available : http://lib.stat.cmu.edu/R/CRAN/.

Bauer, L. (2001). *Morphological productivity*. Cambridge University Press, Cambridge.

Biber, D. (1995). *Dimensions of register variation*. Cambridge University Press, Cambridge.

Gaeta, D. and Ricca, L. (to appear). Productivity in Italian word formation : a variable corpus approach. *Linguistics*.

Geeraerts, D., Grondelaers, S. and Speelman, D. (1999). *Convergentie en divergentie in de Nederlandse woordenschat. Een onderzoek naar kleding- en voetbaltermen*. Meertens Instituut, Amsterdam.

Geerts, G., Haeseryn, W., de Rooij, J. and Van den Toorn, M. C. (1984). *Algemene Nederlandse Spraakkunst*. Wolters-Noordhoff, Groningen.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, vol.(40) : 237–264.

Jelinek, F. and Mercer, R. (1985). Probability distribution estimation for sparse data. *IBM technical disclosure bulletin*, vol( 28) : 2591–2594.

Keune, K., Ernestus, M., Van Hout, R. and Baayen, R. H. (2005). Social, geographical, and register variation in Dutch : From written 'mogelijk' to spoken 'mok'. *Corpus Linguistics and Linguistic Theory*, vol.(1) : 183–223.

Oostdijk, N. (2002). The design of the Spoken Dutch Corpus. In Peters, P., Collins, P. and Smith, A. editors, *New Frontiers of Corpus Research*. Rodopi, Amsterdam : 105–112.

Pauwels, J. L. (1964). Woorden op -atie en -ering in het Nederlands. *Verslagen en Mededelingen van de Koninklijke Vlaamse Academie voor Taal- en Letterkunde* : 205–210.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. Statistics and Computing. Springer, New York.

Plag, I., Dalton-Puffer, C. and Baayen, R. H. (1999). Productivity and register. *Journal of English Language and Linguistics* vol.(3) : 209–228.

Van den Bosch, A. and Daelemans, W. (1999). Memory-based morphological analysis. *In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, ACL'99*, University of Maryland, USA *:* 285-292.