Variation in Dutch: From written MOGELIJK to spoken MOK.

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

[1] *University of Nijmegen, the Netherlands*
[2] *Max Planck Institute for Psycholinguistics, The Netherlands.*

Abstract

In Dutch, high-frequency words with the suffix *-lijk* are often highly reduced in spontaneous unscripted speech. This study addressed socio-geographic variation in the reduction of such words against the backdrop of the variation in their use in written and spoken Dutch. Multivariate analyses of the frequencies with which the words were used in a factorially contrasted set of subcorpora revealed significant variation involving the speaker's country, sex, and education level for spoken Dutch, and involving country and register for written Dutch. Acoustic analyses revealed that Dutch men reduced most often, while Flemish highly educated women reduced least. Two linguistic context effects emerged, one prosodic, and the other pertaining to the flow of information. Words in sentence final position showed less reduction, while words that were better predictable from the preceding word in the sentence (based on mutual information) tended to be reduced more often. The increased probability of reduction for forms that are more predictable in context, combined with the loss of the suffix in the more extremely reduced forms, suggests that high-frequency words in *-lijk* are undergoing a process of erosion that causes them to gravitate towards monomorphemic function words.

1 Introduction

In spontaneous speech words are often pronounced in reduced form (Ernestus, 2000; Johnson, 2004). Some words are reduced to such an extent that an faith-

Email addresses:
karen.keune@mpi.nl,mirjam.ernestus@mpi.nl,r.v.hout@let.ru.nl,baayen@mpi.nl
(Karen Keune,[1] Mirjam Ernestus,[2,1] Roeland van Hout,[1] R. Harald Baayen[1,2]).
[1] This research was supported by a grant from the Netherlands Organization for Scientific Research (NWO). We would like to thank Dirk Geeraerts, Stefan Grondelaers, Dirk Speelman, and Koen Plevoets for stimulating discussion.

ful orthographic transcription would be very different from the orthographic norm. An example from Dutch is the word *mogelijk* ('possible'), which can be pronounced not only as [moxələk] but also as [moxək], [molək], or even as [mok].

Strongly reduced word forms are difficult to interpret without syntactic or semantic context (Ernestus et al., 2002). When speakers of Dutch are presented with the word [mok] in isolation, they find it difficult to assign a meaning to this string of phonemes. It is only when the word is embedded in a sentence that its meaning becomes available. Interestingly, listeners who understood the meaning of [mok] tend to think they heard the full, unreduced form [mox@l@k] (Kemps et al., 2004). A central question in the research on the comprehension of reduced words is what aspects of the linguistic context allow the listener to access the associated semantics.

An important predictor for the degree of reduction in speech production is lexical frequency, as demonstrated by Jurafsky et al. (2001) for function words. The more often a function word is used in speech, the more likely it is to undergo reduction, in line with Zipf's law of abbreviation (Zipf, 1935). Bybee (2001) discussed how frequency of occurrence affects the realization of word final dental plosives in monomorphemic words. **?** observed a negative correlation between frequency and acoustic length for several kinds of derived words in Dutch, including words with the suffix *-lijk*, the suffix in the above example *moge-lijk*. Jurafsky et al. also showed that the degree of reduction is modulated by the extent to which a word is predictable from its context. However, it is currently an open question to what extent the use of reduced forms is codetermined by socio-geographic factors.

Various corpus-based studies have shed light on variation in language use in general. Biber (1988, 1995) identified different varieties of English (and also other languages) by means of factor analyses of the frequencies of a broad range of morphological and syntactic variables. In the domain of literary studies, Burrows (1992a, 1986, 1987, 1992b, 1993a,b) demonstrated regional variation in English narrative, diachronic change in literary texts, and even sex-specific differences in the writing of English historians born before 1850 on the basis of the most common words. Studies in authorship attribution revealed, furthermore, that differences in speech habits can sometimes be traced down to the level of individual language users (Holmes, 1994; Baayen et al., 1996, 2002). Finally, Baayen (1994) and Plag et al. (1999) showed that derivational affixes are used to a different extent in spoken and written registers.

The aim of the present study is to investigate the extent to which the use of words in *-lijk* varies systematically in both written and spoken Dutch. Words in *-lijk* are generally classified as open-class words. However, it is noteworthy that the suffix *-lijk* is hardly productive (Van Marle, 1988), and that many

high-frequency forms are no longer semantically compositional. For instance, *natuur-lijk*, literally 'nature-like', usually means 'of course'. In this study, we will first investigate systematic variation of this unproductive suffix in written Dutch as function of whether a text is written in Flanders or in the Netherlands, and as a function of its register. Second, we explore spoken Dutch as a function of whether a speaker lives in Flanders or in the Netherlands, of the speaker's sex, and the speaker's level of education. Third, we address the question to what extent reduction in the acoustic form of words in *-lijk* is predictable from socio-geographic variables.

In this study, we have made extensive use of multilevel analysis of covariance, a statistical technique that offers two advantages compared to principal components analysis, factor analysis, and correspondence analysis (Lebart et al, 1998). First of all, multilevel modeling allows the researcher to directly assess the significance of predictors, as well as how individual words (or other units of analysis) interact with these predictors. In other words, instead of using both a clustering technique such as principal components analysis and a technique for group separation such as discriminant analysis, we were able to fit a single statistical model to the data that allows us both to trace what predictors are significant, and to visualize their effects. The second advantage of multilevel modeling is that it offers the researcher the possibility to include covariates (such as mutual information) in the model.

2 Written Dutch

For our study of written Dutch, we made use of the CONDIV corpus (Grondelaers et al., 2000). This corpus comprises three kinds of written Dutch: written Dutch from newspapers, written Dutch from USENET, and written Dutch from chat sites. In the present study, we investigated lexical variation in the subcorpus of newspapers. The CONDIV corpus sampled four Flemish newspapers (*De Standaard*, *Het Laatste Nieuws*, *De Gazet van Antwerpen* and *Het Belang van Limburg*) and three Dutch newspapers (*NRC Handelsblad*, *De Telegraaf* and *De Limburger*). These seven newspapers can also be cross-classified according to their register. *De Standaard* and *NRC Handelsblad* are Quality newspapers, aiming at a more educated readership. *Het Laatste Nieuws* and *De Telegraaf* are National newspapers, and *De Gazet van Antwerpen*, *Het Belang van Limburg*, and *De Limburger* are Regional newspapers.

For each of the seven newspapers in the CONDIV corpus, we selected the first 1.5 million words (the size of the smallest newspaper) for further analysis. From these data sets, we selected the 80 most frequent words in *-lijk* (listed in the appendix) that occurred at least once in each of the seven subcorpora,

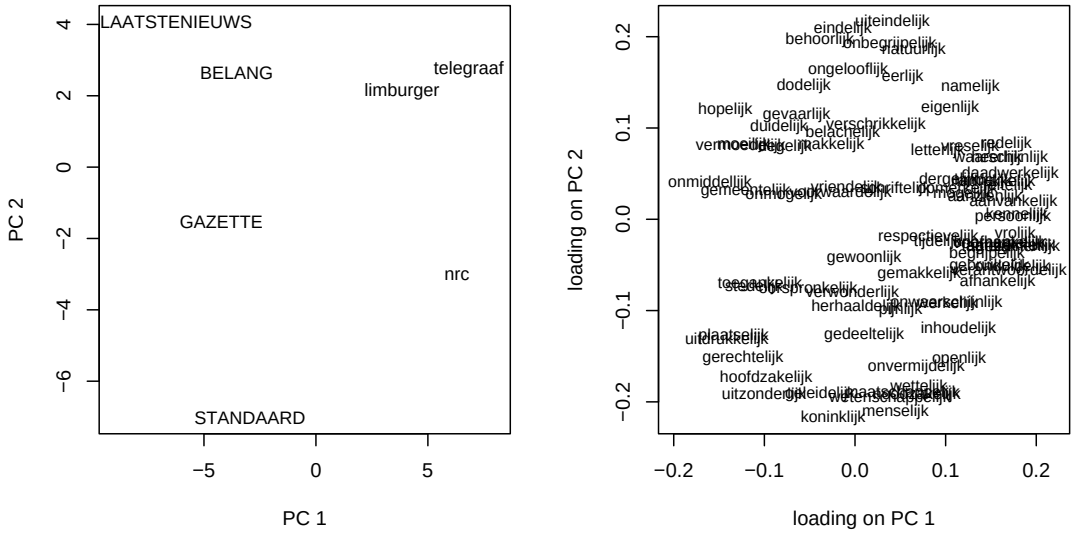

Figure 1. Principal Component Analysis of 80 words ending in *-lijk* in the seven (CONDIV) newspapers. The names of Flemish newspapers are in capital letters.

and registered their frequencies in these subcorpora, which we cross-classified by Country and Register. (Pooling the most common words in each of the subcorpora separately led to a change in only one word.) In this way, we obtained a table with 80 rows (words) and 7 columns (newspapers). One way of looking at these data is that the seven newspapers are represented as 7 points in an 80-dimensional space. This raises the question whether the way in which these seven newspapers are distributed in this space reflects the Registers and Countries of these newspapers.

There are many different statistical techniques for addressing this question, among which principal component analysis, factor analysis, and correspondence analysis are currently the most widely used. Each of these techniques allows the researcher to explore the structure among our newspapers by means of dimension reduction. Figure 1 summarizes the results of a principal component analysis. The left panel plots the newspapers in lexical space by means of the first two principal components. The first principal component (PC1) accounted for 37.3% of the variance, the second (PC2) accounted for 20.1% of the variance. As can be seen in the left panel of Figure 1, these two components reflect the geographic and register differences between the newspapers. First consider PC1. The Flemish newspapers, represented in upper case letters, occur more to the left of the graph, while the Dutch newspapers appear more to the right. In other words, PC1 captures the geographical variation in the use of the 80 high-frequency words in *-lijk* that we sampled. PC2, on the other hand, captures aspects of the register variation. The Quality newspa-

pers (*NRC Handelsblad*, denoted by `nrc` in the plot, and *De Standaard*) appear lower in the plot, while the National newspapers, *Het Laatste Nieuws* and *De Telegraaf* appear at the top of the graph. In the right panel the loadings of the target words on the newspapers is plotted. Words positioned lower in the plot, for instance, have the highest load on *De Standaard*, and are thus most often used in that newspaper.

In order to ascertain to what extent this interpretation is statistically robust, we carried out two tests contrasting the coordinates of the newspapers on the two principal components. A Welch Two Sample t-test contrasting the Flemish and Dutch newspapers with respect to PC1 revealed a highly significant difference ($t(4.16) = -8.47, p = 0.0009$), and a one-way analysis of variance contrasting the three Registers with respect to PC2 also revealed significant differences ($F(2, 4) = 8.07, p = 0.0394$).

Although these tests support the conclusions we drew from the visual inspection of Figure 1, there are a number of questions that this exploratory technique does not answer. One of these questions concerns the possibility of an interaction between Country and Register. Do these two factors work independently, or might the effect of one of these factors be different depending on the value of the other factor? Second, are these geographic and register differences supported in the same way by each of our 80 words? It might be the case that the main effects uncovered by the principal components analysis are supported only by specific subsets of words. More technically, we would like to be able to ascertain whether there are interactions between the words and Register and Country. We therefore analyzed these data in more detail using multilevel regression modeling.

Multilevel modeling (Pinheiro and Bates, 2000) is a regression technique developed to deal specifically with data combining fixed and random effects. Factors are described as 'fixed' when the levels of that factor exhaust all possible levels. An example of a fixed effect in the present data is Country: the Netherlands and Flanders are the only two European countries in which Dutch is spoken, there are no other conceivable levels of this factor that we have not sampled. By contrast, the words in our data set constitute a 'random' effect: these words are sampled from a larger population of words in *-lijk*, and we would like to know whether the patterns observed in the data would generalize to the whole class of words in *-lijk*. In the model that we fit to these data, we therefore included Word as a random factor, it is the main grouping factor in the analyses to follow. Mixed effects models deal with the distinction between fixed and random effects in a more principled way than do traditional linear models, and, more importantly, they provide more precise estimates of the random effects (in this study, improved estimates of the effects of the individual words). In addition, these by-word adjustments are easier to extract and inspect than with standard or general linear models (Quené & Van den

Bergh, 2004; Baayen, 2004).

.

Recall that we have 7 observations for each word, one frequency count for each newspaper. One way of looking at what multilevel modeling does is to build informed models for each of the individual words. The individual models are informed in the sense that they are constructed against the background of what is known about the behavior of all the other words in the sample.

A multilevel model fit to the logarithmically transformed frequencies of the 80 words in *-lijk* in the seven newspapers (using a stepwise model selection procedure), with Word as grouping factor, revealed a significant (fixed) effect for Country ($F(1, 463) = 9.3067, p = 0.0024$), a marginally significant (fixed) effect for Register ($F(2, 463) = 2.4592, p = 0.0866$), and a significant interaction of Country by Register ($F(2, 463) = 16.1930, p < 0.0001$). The frequencies of words in *-lijk* tended to be lower in Flanders compared to the Netherlands. In both countries, words in *-lijk* were used most frequently in the Quality newspaper. Furthermore, in Flanders words in *-lijk* were used significantly less often in the National newspaper than in the Quality newspaper. Conversely, in the Netherlands words in *-lijk* were used significantly less often in the Regional newspaper than in the Quality newspaper. This model provides further support for the general patterns discovered by the principal components analysis. However, it also provides a correction by uncovering an interaction of Country by Register. In addition, the multilevel model points not only to a difference between Flanders and The Netherlands with respect to the use of words in *-lijk*, but also discloses that, apparently, words in *-lijk* are used slightly more often in the Netherlands.

In multilevel modeling, it is also possible to investigate whether there are interactions between the fixed effects and the main grouping factor Word. We observed significant interactions involving Word both for Country and for Register ($p < 0.0001$ and $p < 0.0023$, likelihood ratio tests). There are two further technical details concerning this model. First, we removed outliers from the data set, i.e., data points with standardized residuals with an absolute value exceeding 2 standard deviation units (see Chatterjee et al. 2000 for further details on the removal of outliers in multiple regression). In the present model this led to the removal of 12 data points (2.1% of the 560 data points). Second, we added an extra parameter to the model in order to remove the heteroscedasticity visible in the plot of the standardized residuals against the fitted values. This extra parameter (for an exponential variance function, see Pinheiro & Bates 2000, 211-213) was also justified by a likelihood ratio test ($p < 0.0001$).

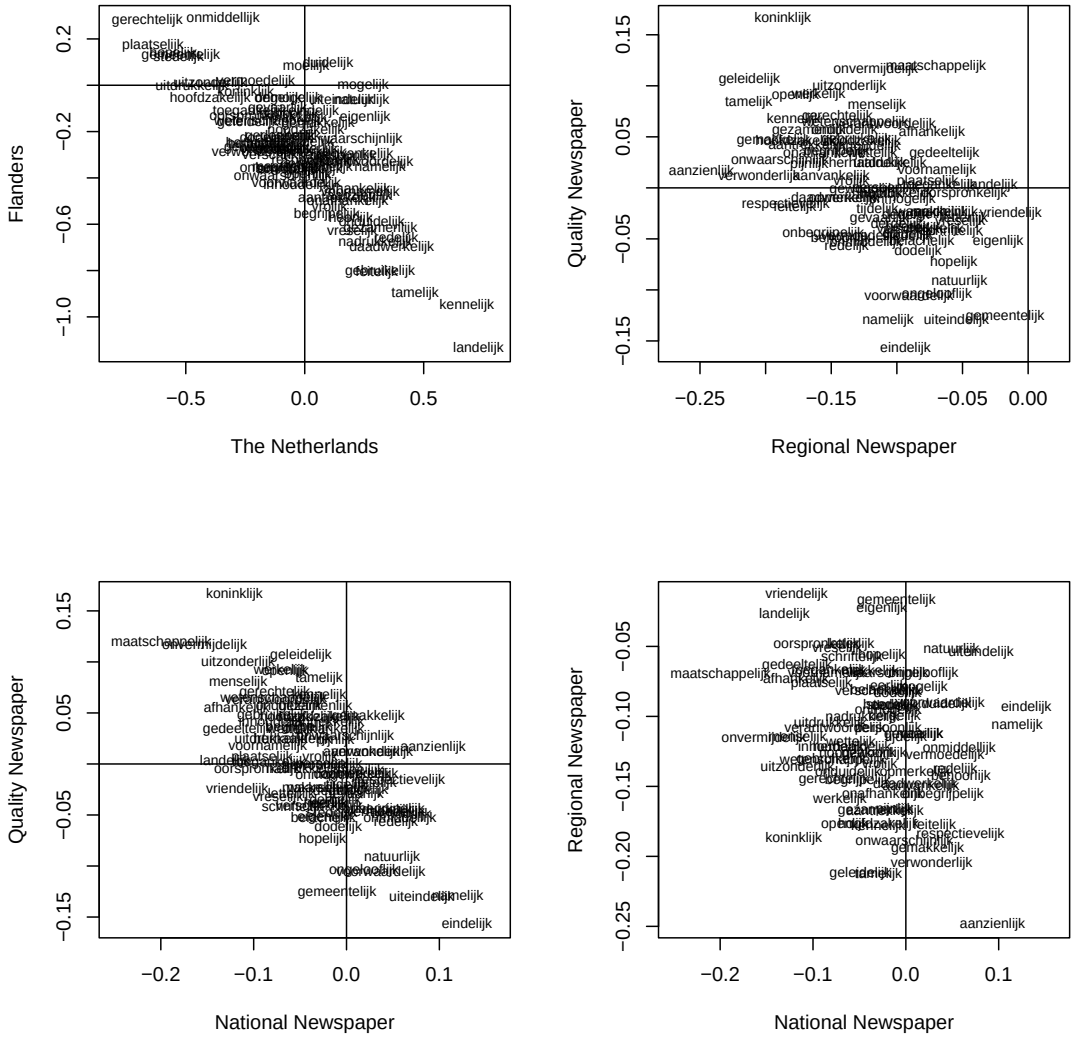

Figure 2. By Word adjustments for Country and Register in a multilevel model for 80 selected words ending in *-lijk* from the seven CONDIV newspapers.

Figure 2 provides a visual aid to understanding the interactions involving Word. The upper left panel shows the interaction of Word by Country. Recall that we observed a main effect for Country, with words in *-lijk* being used more frequently in the Netherlands. The interaction of Country by Word shows that this effect is not equally strong for all words. The horizontal axis of the upper left panel shows the by-word adjustments that need to be made in order to make the predictions for the frequencies of the words as used in the Netherlands more precise. The vertical axis does the same for the predictions pertaining to the Flemish frequencies. Positive values indicate that a word is used more often than the average word in *-lijk* in the country associated with the axis. In other words, the further to the right a word is positioned, the more

frequently it is used in the Netherlands. The higher a target is positioned, the more frequently it is used in Flanders. The words, *landelijk* ('national') and *kennelijk* ('apparently'), for instance, are used more often in the Netherlands than in Flanders, while *onmiddellijk* ('immediately'), and *gerechtelijk* ('judicial') are used more often in Flanders. [2] We listed the coordinates of all words in Figure 2 as well as the coordinates of all words in the following figures in Appendix B.

A closer inspection of this plot and the corresponding table of by-word adjustments suggests that the locatives *gemeentelijk* ('municipal'), *plaatselijk* ('local') and *stedelijk* ('urban') are used more frequently in Flanders while the locative *landelijk* ('national') is used more frequently in the Netherlands. Moreover, there are two near-synonyms for *explicit(ly)* that show differential use across the two countries: *Uitdrukkelijk* is typically Flemish and *nadrukkelijk* is typically Dutch.

The remaining three panels of Figure 2 plot the register variation for words in *-lijk*. The upper right panel, for instance, shows the variation of use of words in *-lijk* for Regional compared to Quality newspapers. For example, the word *gemeentelijk* ('municipal') appears more frequently in Regional newspapers, and less frequently in Quality newspapers. Words typical for the Quality newspapers are, among others, *koninklijk* ('royal'), *onvermijdelijk* ('inevitable'), and *geleidelijk* ('gradual'). A word typical for the Regional newspapers is *gemeentelijk* ('municipal'). Note that most words appear more frequently in the Quality newspapers than in the Regional newspapers.

The question that arises at this point is whether the geographic and register variation in the use of words in *-lijk* is specific to these particular complex words, or whether this variation is also reflected in the use of other aspects of lexis and grammar. In other words, we need an independent and established method for tracing variation in other parts of grammar and lexis in order have a benchmark with which the present results can be compared.

The benchmark that we selected is the stylometric technique developed by Burrows (1992a, 1993a). Burrows showed that differences in speech habits of individual language users are reflected in their use of the most common word types. The most common words typically include function words (determiners,

[2] Multilevel models only specify whether an interaction involving the main grouping factor (Word in the present example) is significant, but do not provide means for comparing the significance of differences involving individual words. Questions such as whether a given word occurs significantly more often in Flanders or in the Netherlands require independent statistical tests, for instance, tests based on contingency tables such as Fisher's exact test of independence. Note that such independent tests are justified only in the present framework for comparisons for which significant interactions with the main grouping factor have been observed.

pronouns, conjunctions, auxiliaries) as well as some common adverbs. Differences in the use of the most common words tend to represent differences in syntactic habits (Baayen et al., 1996). Content words are usually excluded from the list of most common words in stylometric studies, in order to avoid clustering based on topical rather than on structural linguistic features. We applied this state-of-the-art approach from stylometry not at the level of individual speakers but at the aggregate level of groups of speakers defined by socio-geographic variables. We used the same corpus of Dutch and Flemish newspapers, and selected the 80 most common words, excluding 3 content words from this list. These words are listed in the appendix.

A multilevel model fit to the logarithmically transformed frequencies of the 80 most common words that appeared in each of the seven newspapers revealed significant main effects for Country ($F(1, 463) = 41.478, p < 0.0001$), Register ($F(2, 463) = 50.854, p < 0.0001$) and an interaction of Register by Country ($F(2, 463) = 45.168$, $p < 0.0001$). There were no significant differences pertaining to Register within the set of Dutch newspapers. Within the set of Flemish newspapers, the Regional newspapers used the 80 most common words equally often as the Dutch newspapers, in contrast to the Quality newspapers, which used them least often. Similar to the case of the words in *-lijk*, we observed significant interactions between the main grouping factor (Word) and Country as well as Register (both $p < 0.0001$, likelihood ratio tests). This model was obtained after removing 12 influential outliers (2.1% of the 560 data points). We again used an exponential variance function in order to remove heteroscedasticity visible in the plot of the standardized residuals against the fitted values. As before, this involved adding an additional parameter to the model, which was justified by a likelihood ratio test ($p < 0.0001$).

When we collapse over the different registers, we find that the most common words in the present study are used less often in Flanders than in the Netherlands. This is probably due to the selection of only the 80 most frequent common words for analysis. The dialects of Flanders are characterized by a much greater variety of forms than those in the Netherlands, especially in the pronominal system. Furthermore, the standard language in Flanders is more divorced from the language varieties used in informal communicative situations compared to the Netherlands. We suspect that some dialectal variants were used in the Flemish materials along with the standard forms. If so, the standard forms were used somewhat less frequently than in the corresponding Dutch texts.

Interestingly, the interaction of Country by Register shows that this difference between Dutch and Flemish emerges most markedly for the Quality Belgian newspaper *De Standaard*. Since this journal is known to use a rather formal style, the low frequency of most common words in this journal cannot be ascribed to the presence of dialectal forms. It is more likely that this difference

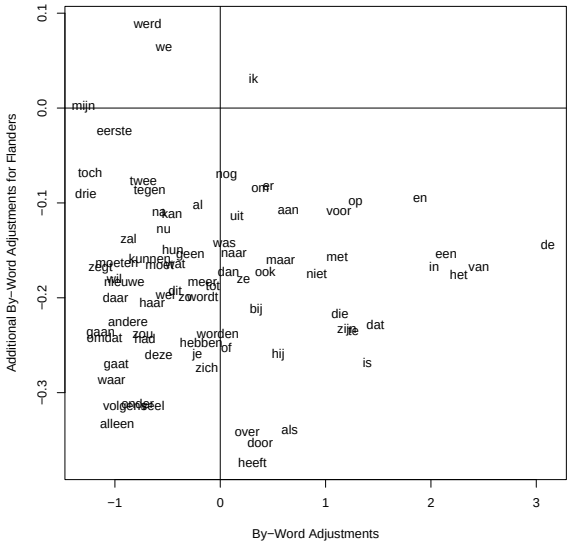

Figure 3. By Word adjustments for Country in a multilevel model for 80 selected most common word types from the seven CONDIV newspapers.

suggests that the journalists writing for this newspaper use more content words in their articles than journalists writing for the other newspapers, which leads to a higher information density.

Figure 3 illustrates the interaction of Word by Country. The x-axis shows the by Word adjustments of the relative frequency necessary to obtain an accurate estimate of the relative frequency of each word as used in the Netherlands. The y-axis shows the extra by Word adjustments needed to obtain the relative frequency of the words as used in Flanders. The word *ik* ('I'), for instance, is used more frequently in Flanders than the average most common word in the data set, as shown by its high value on the y-axis.

The way in which the different most common words are positioned suggests that the first person pronouns (*we* 'we', *ik* 'I') are used more often in Flanders, while third person pronouns are used more frequently in the Netherlands (*hij* 'he', *zij* and *ze* 'she', *zich* 'oneself'). Note that only three words are used more frequently in Flanders than in the Netherlands.

When we compare the socio-geographic variation observed for words in *-lijk* with the variation as indicated by the most common words, we find both similarities and differences. Both sets of words emerged as carriers of socio-geographic differentiation. Furthermore, both the most common words and the words in *-lijk* were used somewhat less often in Flanders. With respect to register, however, the two sets led to different results: the most common words were atypical for quality papers in Flanders, while the words in *-lijk* were more characteristic of quality newspapers in both regions.

3 Spoken Dutch

Next we explored effects of socio-geographic variation on the frequency with which words in *-lijk* and most common words are used in spoken Dutch.

We made use of the Corpus of Spoken Dutch (CGN) (Oostdijk, 2002). This corpus contains approximately 8.9 million words of spoken Dutch, sampled from a wide range of registers. In order to maximize the contrast between written and spoken Dutch, we focused on the subcorpora containing recordings of spontaneous speech. The CGN comprises two categories of spontaneous Dutch: face-to-face conversations and telephone dialogues, in all 4,7 million words. The CGN provides detailed information about the different speakers including the country in which they live, their education level, and their sex. This made it possible for us to not only investigate the effects of Country (the Netherlands versus Flanders), but also the effects of Education (high (attended bachelor or master education) versus non high education level), and Sex (men versus women).

We created eight subcorpora according to a 2x2x2 factorial design with as factors Country, Sex, and Education. These subcorpora differed substantially in size, ranging from 189,000 words (for Flemish male speakers with a non high education level) to 1,200,000 words (for Dutch female speakers with a high education level). We then selected all words in *-lijk* that appeared at least once in each of these eight subcorpora (32 words, see appendix), and we calculated their relative frequencies in each subcorpus. These relative frequencies were the dependent variable in a multilevel model with Word as main grouping factor and Country, Sex, and Education as predictors.

In contrast to the results of our study of words in *-lijk* in written Dutch, the model fit to the logarithmically transformed relative frequencies revealed no significant main effect for Country. There was also no significant main effect for Sex. However, speakers with a higher education level tended to use words in *-lijk* more often than speakers with lower education levels ($F(1, 218) = 4.0514, p = 0.0454$). The main effect of Education in spoken Dutch mirrors the greater use of *-lijk* in the Quality newspapers as compared to the National newspapers, with as main difference that in spoken Dutch this simple main effect is not modulated further by an interaction with Country.

Furthermore, we observed significant interactions of Word by Country, Word by Sex, and Word by Education ($p < 0.001$, likelihood ratio tests). This model was obtained after removing six influential outliers (2.3% of the 256 data points). The heteroscedasticity visible in the plot of the standardized residuals against the fitted values was again brought under control with an additional parameter for the variance function, justified by a likelihood ratio test ($p <$

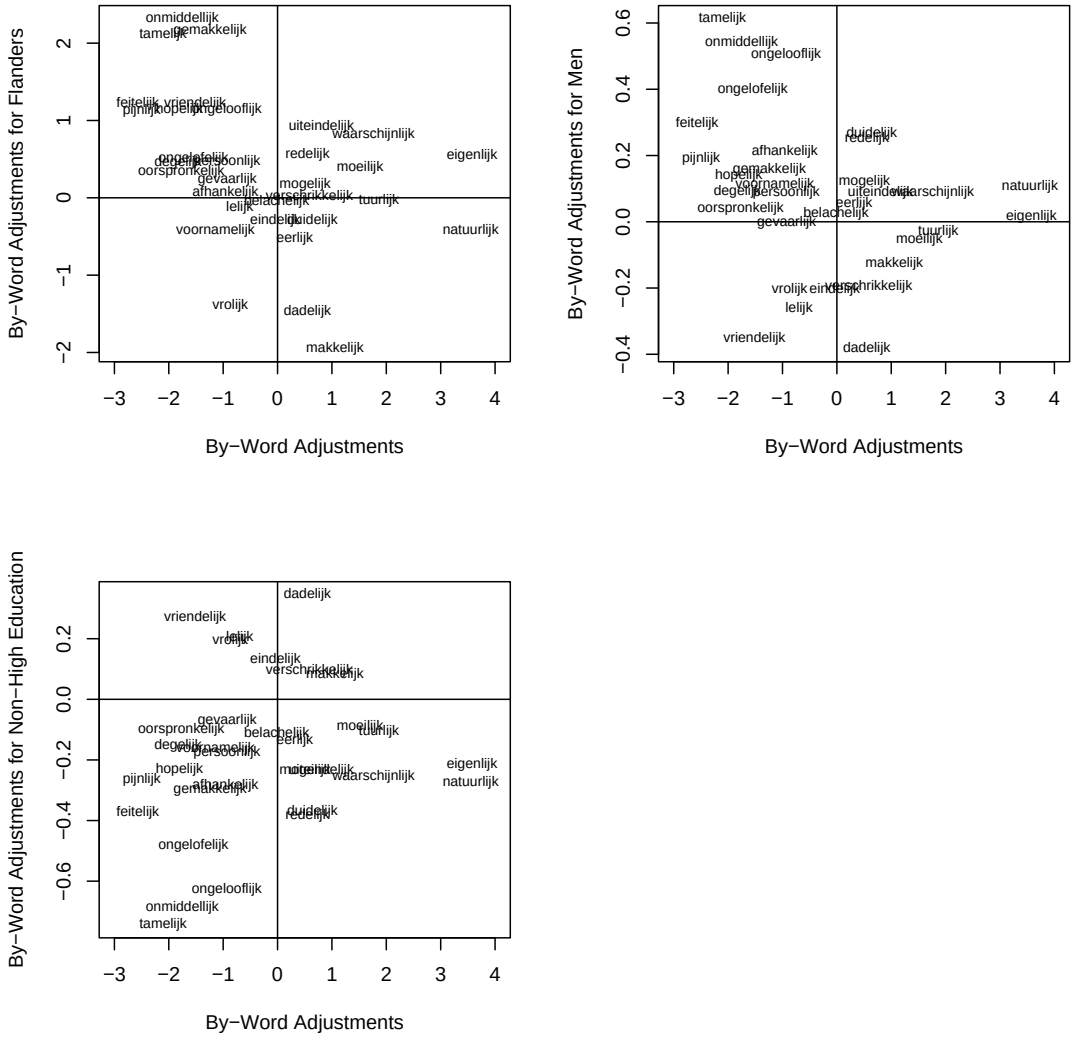

Figure 4. By Word adjustments for Country, Sex, and Education in a multilevel model for 32 selected words ending in *-lijk* from the eight, factorially designed, subcorpora of the spontaneous conversations in the CGN.

0.0001).

Figure 4 illustrates these interactions between Word and Country, Sex, and Education. All x-axes show the by-words adjustments necessary to obtain the relative word frequencies for Dutch women with a high education level. The y-axis of the upper left panel shows the extra adjustment of the relative frequency required for each word to obtain an accurate estimate for the relative frequency in Flanders (Flemish highly educated women). The words *eigenlijk* ('actually') and *natuurlijk* ('of course'), for instance, are used very frequently in the Netherlands. However, *eigenlijk* is used even more frequently in Flanders, while *natuurlijk* is used somewhat less frequently in Flanders.

In addition, this panel shows that *onmiddellijk, gemakkelijk* and *tamelijk* ('immediately', 'easily', 'somewhat') are typical for Flanders, while *vrolijk, dadelijk* and *makkelijk* ('happy', 'immediately', 'easily') are typical for the Netherlands. Interestingly, *onmiddellijk* and *dadelijk* are (near) synonyms for 'immediately', and *gemakkelijk* and *makkelijk* are variants of 'easily'. It is standard practice in sociolinguistics to investigate linguistic variation in time and space by means of pairs of expressions that differ in one dimension only. At the lexical level, this implies that only pairs such as *gemakkelijk/makkelijk* and *dadelijk/onmiddellijk* would be used to probe sociolinguistic variation at the lexical level. What the methodology explored in the present study allows us to observe is that such matched pairs of words indeed are strong carriers of variation, but that there are other, non-matched words such as *tamelijk* ('somewhat') and *vrolijk* ('happy') that are also involved in this geographical opposition.

The upper right panel shows the by-word effects for Sex, with on the y-axis the adjustment for frequency of use for men (Dutch highly educated men). The near synonyms of 'immediately', *onmiddellijk* (men) and *dadelijk* (women) are clear markers for the two sexes, but, as before, there are also other, non-synonymous markers, such as *tamelijk* and *ongelooflijk* ('somewhat', 'unbelievable', more typical for men) versus *vriendelijk* and *lelijk* ('friendly' and 'ugly', more typical for women).

The lower left panel illustrates the required adjustment of the relative frequency of the different words for non highly educated speakers (Dutch non-highly educated women). It shows that the synonyms of 'immediately' also differentiate between the education level of speakers, together with *vriendelijk, lelijk, vrolijk* ('friendly', 'ugly', 'happy' for non high education) and *tamelijk, ongelooflijk* ('somewhat', 'unbelievable' for high education).

As in the analyses of written Dutch, we investigated whether the socio-geographic variation in the use of *-lijk* is also reflected in the use of the most common words. A multilevel model fit to relative frequencies, raised to the power of 0.25 [3], of the 80 most common words occurring in all eight subcorpora (see appendix) revealed only one significant main effect: Men tended to use the 80 most common words less often than women ($F(1, 551) = 14.759, p = 0.0001$). This suggests that the speech of men is characterized by a slightly higher information density compared to women. (This higher information density may be due to more intensive use of less common non-content words, but also to the use of more content words.) We observed significant interactions between the main grouping factor Word, and Country, Sex, and Education (all $p < 0.0001$, likelihood ratio tests). This model was obtained after the removal of eight

[3] This transformation brought the distribution of relative frequencies more in line with the normality assumptions underlying linear regression.

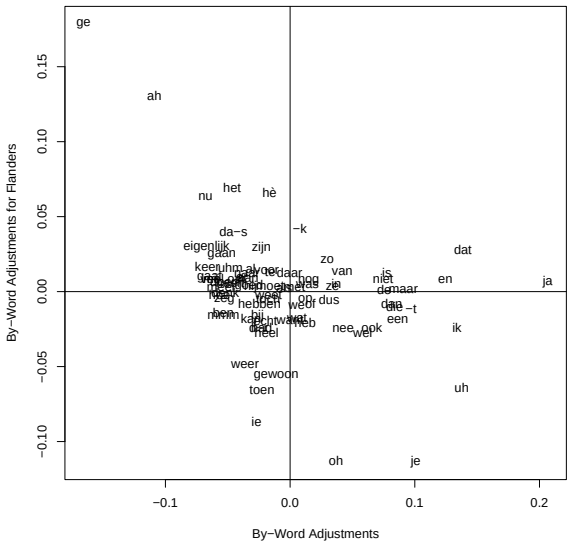

Figure 5. By Word adjustments for Country in a multilevel model for the 80 most common words from the eight, factorially designed, subcorpora of the spontaneous conversations in the CGN.

influential outliers (1.3% of the 640 data points).

The interaction of Word by Country is illustrated in Figure 5. Again, the x-axis shows the by Word adjustments to obtain the relative frequency of the different words as used in the Netherlands, and the y-axis shows the extra adjustment of the relative frequency for each word to obtain the relative frequency of the word as used in Flanders. The near-synonyms *ge* ('you', Flanders) and *je* ('you', the Netherlands) are clear markers of the differences in lexis across countries. Furthermore, the discourse marker *ah* appears to be preferred by Flemish speakers, and *oh* by Dutch speakers.

Summing up, in both spoken and written Dutch, and for both words in - *lijk* and most common words, all predictors interacted with Word. Furthermore, there were differences in the main effects. For words in *-lijk*, we observed that speakers with a higher education level used these words more often, and so did the Dutch Quality newspaper. The selected most common words in spoken Dutch were used less frequently by men than by women. Furthermore, the most common words selected from written texts were used more frequently in the Netherlands compared to Flanders, and in Flanders differentiated between the different kinds of newspapers.

4 Variation in the reduction of -lijk

The preceding analyses of *-lijk* in spoken Dutch proceeded on the basis of the orthographic transcriptions of spontaneous conversations. These analyses glossed over a property of these words that is a potential carrier of socio-geographic differences, namely, the extent to which these words are reduced acoustically in casual speech.

In order to explore this potential socio-geographic stratification of acoustic reduction, we selected those words in *-lijk* that occurred more than 75 times in the subcorpus of spontaneous Dutch from the set of 32 words in *-lijk* examined above. For these 24 words, we aimed at randomly selecting the acoustic signal for ten occurrences in each of the eight cells of the design obtained by factorially contrasting Country, Sex, and Education. In roughly one third of the cases it turned out to be impossible to obtain even ten occurrences, either because of data sparseness or because of a variety of problems with the acoustic signal itself. Instead of the desired 80 tokens for each of the 24 selected words the mean number of observations for a word in our design was 64.3, the median was 64 and the range was 43 to 80. The total number of observations was 1543.

A broad phonological transcription, made by one transcriber, for each of these 1543 sound files served as the basis for assignment to one of three levels of Reduction: No Reduction, Medium Reduction, and High Reduction. Words were classified as having No Reduction either when both the suffix and the stem were fully preserved [moxələk] (mogelijk, 'possible'), [tyrlək] (tuurlijk, 'of course') or when the suffix *-elijk* was reduced to *lijk* and the stem was fully preserved *[moxlək]* (*mogelijk*). Reduction of the suffix from *-elijk* to *-lijk* was not classified as reduction for two reasons. The first reason was that both *-elijk* and *-lijk* are allomorphs of the same suffix. Which allomorph is used, depends only on the phonemes preceding the suffix. The second reason was that it was often hard to ascertain whether or not the schwa was still present in the suffix. Words were classified as having Medium Reduction when the /l/ from the suffix or when consonants from the coda of the stem were not present *[molək]*, (*mogelijk*), *[moxək]*, (*mogelijk*), *[ɛɪnlək]* (*eindelijk*), *[ɛɪdlək]* (*eindelijk, 'finally'*). If the coda of the stem had more than one consonant, one of these consonants and the /l/ of the suffix could be absent: *[ɛɪnək], [ɛɪdək], [ɛɪlək]* (all forms of *eindelijk*). Words were classified as having High Reduction either when the suffix was completely integrated with the stem, with the final /k/ of the suffix becoming the coda of the stem *[mok]* (*mogelijk*), *[moxk]* (*mogelijk*), or when the suffix had disappeared completely *[mo]* (*mogelijk*).

Of the 24 initially selected words, represented by 1543 tokens, only 14 words appeared in a Medium or High Reduced form: *afhankelijk* ('dependent'), *da-*

delijk ('immediately'), *duidelijk* ('clear'), *eerlijk* ('honest/fair'), *eigenlijk* ('actually'), *eindelijk* ('finally'), *moeilijk* ('difficult'), *mogelijk* ('possible'), *natuurlijk* ('of course'), *persoonlijk* ('personal'), *tuurlijk* ('of course'), *uiteindelijk* ('finally'), *vriendelijk* ('friendly'), and *waarschijnlijk* ('probably'), in all 946 word tokens. In order to provide some validation for the three categories of reduction and the initial assignment of the word tokens to these categories, a second judge also listened to each of these 946 words and assigned them to one of the three reduction categories. For 19 tokens the new assignment deviated from the original one. A third judge determined the final assignment for these word tokens. We calculated two statistics for each of the words: the relative frequency of the word in a given subcorpus, and the mutual information (Church and Hanks, 1990; Gregory et al., 1999) of the word and the word preceding it, which estimates the predictability of a word given the preceding word in the sentence. (For words with a frequency less than 11 in a subcorpus, the mutual information was set to zero in order to avoid excessively high and uninformative mutual information values.) Finally, we registered whether a token occurred in the final or in a non-final position in the sentence.

In the preceding statistical analyses we used multilevel models with Word as main grouping factor. By modeling Word as a random effect, the results of the statistical analyses generalized to the population of (higher frequency) words from which we sampled our materials. Since we have only 14 words ending in *-lijk* in the present data set, and since these 14 words are in no way a random sample, we opted for analyzing Word as a fixed effect in the analyses to follow.

Six of the 14 words were characterized by only two levels of reduction (High Reduction versus No Reduction). For eight words, all three levels of reduction were attested in colloquial Dutch. In what follows, we analyzed the log odds ratio of the number of words with No Reduction to the number of words with High or Medium reduction, using logistic regression (Harrell, 2001).

A logistic simple main-effects model of covariance fitted to the 946 data points (using a stepwise model selection procedure) revealed significant effects for Country ($X^2_{(1)} = 13.15, p = 0.0003$), Sex ($X^2_{(1)} = 7.35, p = 0.0067$), Position ($X^2_{(1)} = 6.69, p = 0.0097$), Mutual Information ($X^2_{(1)} = 7.83, p = 0.0051$), and Word ($X^2_{(13)} = 235.10, p < 0.0001$). When we allowed two-way interactions into the model, interactions emerged of Country by Education ($X^2_{(1)} = 6.09, p = 0.0136$) and of Country by Word ($X^2_{(13)} = 26.12, p = 0.0164$). Due to the latter interaction, the main effect of Country, which revealed that speakers in Flanders reduced less than speakers in the Netherlands, was no longer significant. Thus, it appeared that words in *lijk* are overall more often reduced by Dutch speakers, but that this is not the case for all the individual words. The coefficients of this model, with the exception of those involving the interaction of Country by Word, are summarized in Table 1. The generalized R^2 index for this model was 0.568, and Somer's D_{xy} was 0.778.

Table 1
Coefficients in the logistic regression model (with dummy contrast coding) for the suffix reduction data. The intercept represents the log odds ratio of non-reduced to reduced words for Dutch highly educated women. The coefficients show the change in the log odds ratio accompanying the change from, e.g., women to men. See also Figure 6.

	Coefficient	Wald Z	p
Intercept	1.94	3.84	0.0001
Country: Flanders	0.58	0.87	0.3862
Sex: Male	-0.49	-2.67	0.0075
Position: Non-Final	-0.78	-3.10	0.0019
Education: not High	0.32	1.36	0.1726
Mutual Information	-0.11	-2.73	0.0064
Country: Flanders, & Education: Not High	-0.94	-2.47	0.0136

The partial effects of the main predictors in the model are illustrated in Figure 6. The upper left panel graphs the observed proportions of reduced forms for men and women: Men reduce more often than women. This may be due to the higher speech rate of men compared to women (Verhoeven et al., 2004). The upper right panel illustrates that words were less likely to be reduced in sentence final position. The lengthening of words in phrase-final position has been found to be a parsing cue for the listener (e.g., Scott, 1982). Reducing words in *-lijk* in phrase-final position would result in the absence of a useful perceptual cue, and apparently is avoided. The interaction of Country by Education is illustrated in the lower left panel. The factor Education is predictive only for Flanders: Flemish speakers with a high education level reduce less than non-highly educated Flemish speakers. As can be seen in Table 1, the coefficient for education (0.32), which describes the situation for the Netherlands, was not significant ($p = 0.17$). The lower right panel shows that reduced word forms had a higher mutual information. When words have a reduced information load, their forms can be less distinct as well.

The interaction of Country by Word is summarized in Figure 7. The horizontal axis indicates the adjustment that has to be made to express the amount of reduction of a given word in the Netherlands in relation to the amount of reduction of the word that is least often reduced, namely *moeilijk* ('difficult'), in the Netherlands. The vertical axis shows the adjustments for the amount of reduction of these words in Flanders, again compared to the amount of reduction of *moeilijk* in the Netherlands. The more negative the value on the axes, the greater the likelihood of reduction. Thus, the words in the upper left are relatively often reduced in the Netherlands, but are relatively seldom

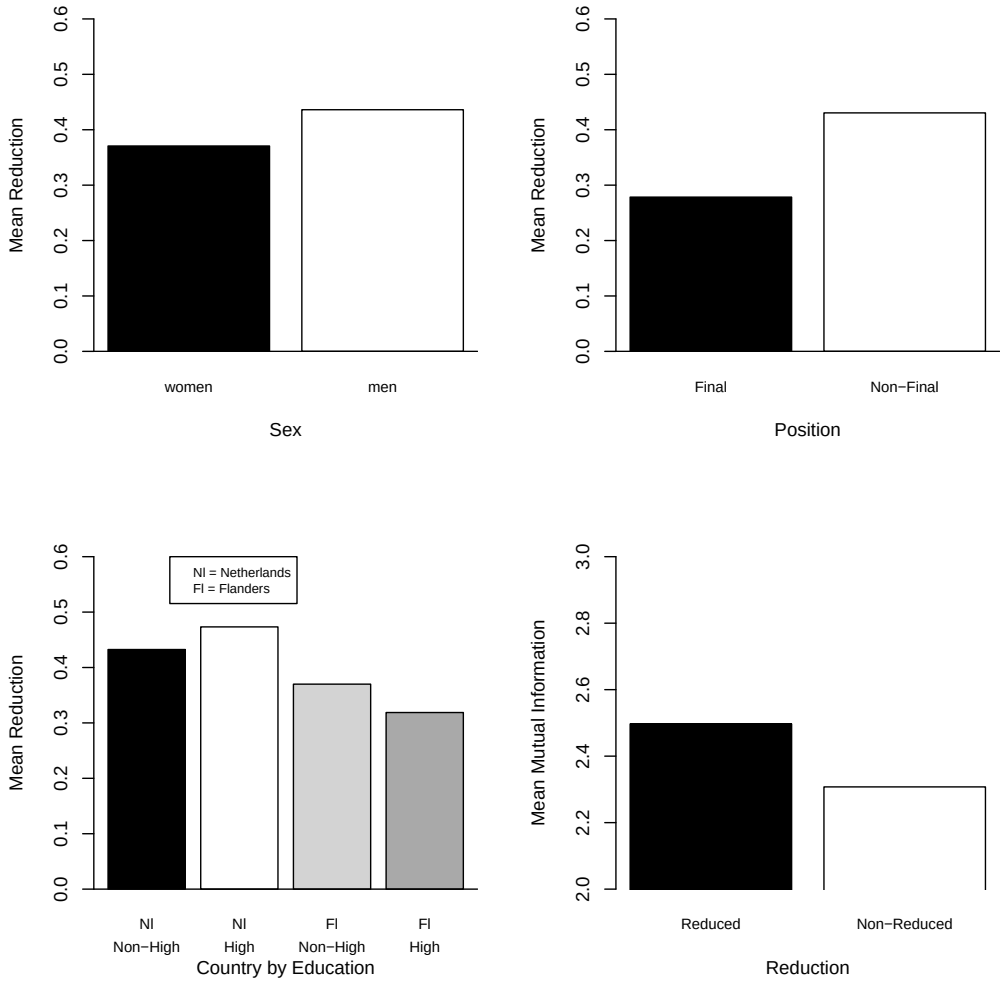

Figure 6. Observed proportion of reduced forms for 14 high-frequency words in *-lijk* broken down for Sex, for Position, and for both Country and Education. The lower right panel plots the mean Mutual Information for the reduced and unreduced forms of 14 high-frequency words in *-lijk*.

reduced in Flanders.

Note that the words *dadelijk* ('immediately'), *uiteindelijk* ('finally') and *tuurlijk* ('of course') differ in their behavior from the other words which are clustered in the lower right of the plot. Nevertheless, the interaction of Word by Country is still significant after the removal of these words. When the word *natuurlijk* is additionally removed, the interaction of Word by Country disappears. So, the behavior of the remaining 10 words is approximately the same in both countries.

Recall that these analyses are based on the log odds ratio of forms without re-

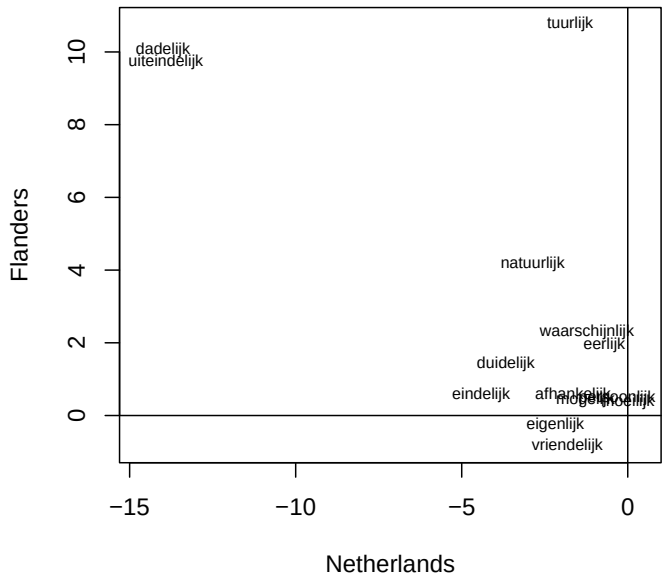

Figure 7. By Word adjustments for Country in a logistic regression model for 14 high-frequency words in *-lijk*.

duction to forms with medium or high reduction. A sub-analysis of the eight words exhibiting three levels of reduction using a proportional odds model (Harrell, 2001) revealed a very similar pattern of results. We also ran an analysis in which we contrasted No or Medium Reduction with High Reduction. In this analysis, Sex and Mutual Information were no longer significant, while the other predictors were retained. This suggests that the effects of Sex and Mutual Information are mainly determined by differences between No and Medium Reduction.

Next, we analyzed the reduction of the vowel in the unstressed, word initial syllable, using the same 946 words we selected to explore variation in the reduction of the suffix. This kind of reduction occurred only in the three words *natuurlijk* ('of course'), *persoonlijk* ('personal'), and *waarschijnlijk* ('probably'). For *natuurlijk* we distinguished /nɑ/ and /na/ from /nə/ and /n/, for *persoonlijk* we distinguished /pə/ from /p/, and for *waarschijnlijk* we distinguished /ʋɑ/ and /ʋa/ from /ʋə/ and /ʋ/. We used the same procedure as before: our point of departure was the abovementioned broad phonological transcription. This transcription was checked by an independent judge, who disagreed in 11 of the 194 cases. For these 11 tokens, a third judge decided their category assignment.

A binary logistic regression model fit to the reduction of the target words *natuurlijk*, *persoonlijk* and *waarschijnlijk* (using a stepwise model selection

procedure) revealed a significant effect for Country ($X^2_{(1)} = 39.59, p < 0.0001$). Dutch speakers reduced the vowel in the unstressed word-initial syllable more often than Flemish speakers. The generalized R^2 index for this model was 0.232, and Somer's D_{xy} was 0.428 .

5 Conclusions

The aim of the present study was to explore the variation in the use of words in *-lijk* in both written and spoken Dutch. First, we investigated variation in written Dutch with respect to the country in which a text is written, and the text's register. Second, we explored spoken Dutch with respect to the speaker's country, level of education, and sex. For spoken Dutch, we investigated in more detail to what degree these socio-geographic factors codetermine the extent to which words in *-lijk* are acoustically reduced.

The methodology that we used for this investigation was inspired by prior stylometric studies on authorial variation (Burrows, 1992a) and studies on register variation (Biber, 1988) which used exploratory multivariate methods such as principal components analysis and factor analysis. We combined insights from these fields with insights from studies investigating the socio-geographical and socio-phonetic forces operating in language (see, e.g., Hay & Cadbury, to appear) in order to increase our understanding of variation in Dutch. Consequently, our study addresses an aggregation level (that of different social groups) that is intermediate between stylometry and authorship attribution (where the linguistic habits of individual authors are of central interest) and register variation (which typically studies texts sampled from a broad range of genres in spoken and written discourse).

Without denying the great value of principal components analysis, factor analysis, and correspondence analysis, we pursued a complementary approach using analysis of variance and covariance of lexical frequencies in factorially contrasted subcorpora. This methodology, which is tailored to our aim of studying the effect of socio-geographic factors on lexical variation, offers several advantages. One such advantage is that it becomes possible to test the significance of the design factors and their interactions with the individual words, without losing the possibility of visualization. Another advantage is that this methodology allows for the possibility of taking covariates into account. Finally, an advantage in relation to standard sociolinguistic practice in which individual controlled variables are studied in isolation, our approach makes it possible to consider a great many potential carriers of sociolinguistic variation simultaneously. This allowed us to trace correlational structure between heterogeneous variables that otherwise remains invisible.

We first studied the variation in the frequency of use of words in *-lijk* in a corpus of Dutch newspapers. We selected all occurrences of 80 high-frequency words in *-lijk* from seven newspapers using a 2 by 3 factorial design. We distinguished between Flemish and Dutch newspapers (Country) and contrasted Quality newspapers (aiming at a more educated readership), National newspapers, and Regional newspapers (Register). In parallel, we conducted a study using the same design based on the 80 most common word types, following Burrows (1986, 1987, 1992ab, 1993ab).

This parallel study was motivated by the hypothesis that variation in the use of *-lijk* is unlikely to be isolated and encapsulated from other dimensions of variation in speech and writing (see Biber, 1988, 1995, for the many correlations into which grammatical markers enter). In order to properly understand the unique contribution of variation in *-lijk* to the linguistic profile of different groups of speakers, we needed a benchmark. Such a benchmark was provided by the covariance structure among the most common words, which tap into the syntactic habits of speakers, and therefore provide a shortcut to the more refined but also far more labor-intensive methods developed by Biber, which are feasible only for well-annotated corpora.

In both analyses, we observed significant and remarkably similar geographic and register differentiation. Apparently, high-frequency words in *-lijk* have a stylometric discriminatory potential that mirrors the well-established stylometric sensitivity of the most common words. Given that words in unproductive *-lijk* constitute a closed-class of words, this is a first way in which high-frequency words in *-lijk* have become to resemble the most common words, which mainly comprise closed-class function words such as conjunctions, pronouns, prepositions, and determiners.

Next, we explored the variation in frequency of use of words in *-lijk* in spoken Dutch. We selected 32 high-frequency words in *-lijk* from the subcorpora of spontaneous face-to-face conversations and telephone dialogues in the CGN, using a factorial design in which we contrasted speakers from Flanders with speakers from the Netherlands, men with women, and highly educated with less educated speakers. As before, we carried out a parallel study using the most common words. This time, we observed a marked difference between the most common words and the words in *-lijk*. Speakers with a higher education level tended to use words in *-lijk* more often. For the Netherlands (but not for Flanders), this mirrors the finding that the quality newspaper made more intensive use of this suffix as well. The analysis of the most common words, by contrast, suggested that men made less use of the most common words compared to women, suggesting the possibility of a slightly higher information density (carried by less frequent closed-class words or even by full-fledged content words) for men. In other words, the comparison with the benchmark for grammatical variation revealed that in spoken Dutch, unlike in written

Dutch, words in *-lijk* tap into an independent source of variation. Furthermore, we also observed significant differences in how individual most common words as well as individual words in *-lijk* were used by men and women in the two countries as a function of their education level.

Finally, we investigated the socio-geographic variation in the degrees of reduction of words in *-lijk*. This kind of research has become possible only recently, thanks to the development of large speech corpora with not only orthographic transcriptions but also the acoustic signal. Corpora such as the corpus of New Zealand English (Schreier et al., 2003; Gordon et al., forthcoming; Hay and Sudbury, to appear) and also the corpus of spoken Dutch offer the possibility of detailed analyses of the variation in acoustic forms across sociolinguistic and stylistic dimensions. The corpus of spoken Dutch was just large enough to allow us to retain our factorial methodology, although it left us with only 14 words ending in *-lijk* (evidencing reduction) that occurred sufficiently often in the different subcorpora defined by crossing Country, Sex, and Education. Two transcribers classified the degree of reduction for a total of 946 tokens of these 14 words. We considered two kinds of reduction, one primarily affecting the suffix, the other affecting the vowel in the word initial syllable. Both analyses showed that in Flanders speakers reduce less than in the Netherlands, which ties in with the more formal status of standard Dutch in Flanders. The reduction involving the suffix was more prominent for men compared to women. Moreover, highly educated Flemish speakers used fewer reduced forms than did less highly educated Flemish speakers. Finally, there were significant differences in the extent to which individual words underwent reduction that we could trace back to the speaker's home country. For instance, *dadelijk* ('deedly', i.e., 'immediately') and *uiteindelijk* ('end-ly', i.e., 'finally') are words that undergo reduction more often in the Netherlands than in Flanders. The degree of reduction is possibly influenced by speech rate. The higher the speech rate is, the more often reduction occurs. This assumption is strengthened by previous research in which, comparable to our results for reduction, it appeared that Dutch men have the highest speech rate, while Flemish women have the lowest (Verhoeven et al., 2004).

In addition to these socio-geographic factors, the degree of reduction was significantly codetermined by two linguistic factors: the word's position in the sentence, and the extent to which the word is predictable from its context. With respect to the word's position in the sentence, we found that words in *-lijk* that occurred in sentence-final position revealed little reduction. This is as expected given that words in sentence final position are often lengthened (e.g., Fougeron and Keating, 1997; Cambier-Langeveld, 2000; Pluymaekers et al., submitted).

We used the mutual information measure to gauge contextual predictivity. Words in *-lijk* with a high mutual information (Manning and Schutze, 1999),

i.e., that exhibited a high degree of predictability from the preceding word, revealed more reduction. As the information load of a word in *-lijk* decreases, speakers fall back gestural scores that require less articulatory effort in production (see cf. Bybee, 2005).

The overall pattern in our data suggests that reduced high-frequency forms in *-lijk*, such as monosyllabic *[tyk]* (for *natuurlijk*, 'of course'), *[mok]* (for *mogelijk*, 'possible') and *[ɛɪk]* (for *eigenlijk*, 'actually') are becoming more similar to the most common words, not only in that they are markers of register and of socio-geographic origin, as observed above, but also in their loss of morphological structure, as witnessed by their lack of semantic compositionality and the erosion (Heine and Kuteva, 2005) of their phonological form.

Interestingly, the position in the sentence and mutual information are contextual predictors that did not interact with the socio-geographic variables. This is reminiscent of the finding of Bresnan et al. (2005) that the formal syntactic and semantic properties governing the dative alternation in English do not change across modality (spoken versus written English), verb sense, and speaker. This suggests that there are robust fundamental linguistic principles that operate in the same way across register and different socio-geographic speech communities. Possibly, phrase-final lengthening and information load belong to the set of these fundamental principles. Further research addressing acoustic reduction for other kinds of complex words is required here.

What the present results clearly show is that for a full explanation of acoustic reduction socio-geographic factors need to be taken into account. Although articulatory explanations (such as offered in Browman and Goldstein, 1992; Ernestus, 2000) increase our insight in the path of acoustic erosion, they do not predict when speakers actually use reduced forms and when they stick with the unreduced forms. We have shown that some headway in predicting degrees of reduction can be made by taking socio-geographic factors and contextual linguistic factors into account.

For *[mok]*, *[ɛɪk]* and *[tyk]* a large series of different forms exist side by side. The unreduced long, morphologically complex, but semantically opaque forms are predominant in the written language, and shape modern speakers' awareness of these words. The reduced, monosyllabic forms are typically found in spontaneous spoken Dutch, without speakers realizing that what they actually say diverges from the written norms (Kemps et al., 2004). Reduced forms challenge models of speech comprehension, which are based on the assumption that words have a single canonical form (Norris, 1994). The present results suggest that listeners might be sensitive to the probability of reduced forms conditional on the socio-geographic and linguistic context in which a word in *-lijk* is uttered, and use this sensitivity to optimize comprehension.

It is well known that morphological rules can cease to be productive. For some affixes, the change from productive to unproductive takes place in a relative short time span, while for others, the change is more gradual (Anshen and Aronoff, 1997, 1999).

However, one of the questions in productivity research is whether there is an absolute distinction between productive and unproductive affixes. Many morphologists believe there is such a distinction (e.g., Schultink, 1961; Anshen and Aronoff, 1999; Bauer, 2001), but there is evidence to the contrary. Baayen (2003) discusses well-formed and contextually natural neologisms in English *-th* (e.g., *coolth*), and a search on the web shows that neologisms in *-lijk* are likewise in use in Dutch. Here are two examples, and many more can be found.

Beschrijf jezelf in 5 woorden: lief agressief aardig **bazelijk** *en gek*
(Describe yourself in 5 words: sweet, aggressive, nice, bossy, and mad)
`www.dreamcommunity.nl/?id=109&account=jEHA2yBJ` (May 2005)

. . . 's ochtends ineens kreeg ik erge spierpijn en werd ik wazig in m'n hoofd, **duizelijk** *en zweverig, buikkrampen en een misselijk gevoel . . .*
(in the morning I suddenly developed muscular pain, I became drowsy in the head, dizzy, woolly, stomach cramps, and I felt sick . . .)
`www.degrotegriepmeting.nl/test/public/index.php?thissection_id=3&request=1150&r=1` (May 2005)

Even though neologisms in *-lijk* have very low probabilities of being coined, there are sufficient numbers of words in *-lijk* in the language for speakers to be able to occasionally generalize *-lijk* to new words. The second example shows that even the blocking force of existing synonyms (*duizelijk* replaces the standard form *duize-lig*) may not prevent new words to be used effectively. From this perspective, the erosion of high-frequency words in *-lijk* is interesting, as this erosion results in words that no longer contribute to the formal similarities that underly this residual productivity of *-lijk*. We expect that as more high-frequency words undergo this process of erosion and become effectively monomorphemic, the residual productivity of *-lijk* will decrease even further. Independent evidence that we are indeed observing a still ongoing process of decreasing productivity and language change is provided by the finding reported by **?** that younger speakers tend to reduce words in *-lijk* to a greater extent then older speakers, and the present finding that reduction is more prominent in the Netherlands than in Flanders (where standard Dutch is used predominantly in more formal contexts).

Anshen, Frank and Mark Aronoff, M.
1997 Morphology in real time. In Booij, Geert. E. and Jaap van Marle (eds.), *Yearbook of Morphology.* Dordrecht: Kluwer Academic Publishers, 9-12.

Anshen, Frank and Mark Aronoff
1999 Using dictionaries to study the mental lexicon. *Brain and Language* 68, 16-26.

Baayen, R. Harald
2003 Probabilistic approaches to morphology. In Bod, Rens, Jennifer Hay and Stefanie Jannedy (eds.), *Probabilistic Linguistics.* The MIT Press, 229-287.

Baayen, R. Harald
1994 Derivational productivity and text typology. *Journal of Quantitative Linguistics* 1, 16-34.

Baayen, R. Harald
2004 Statistics in psycholinguistics: A critique of some current gold standards. *Mental Lexicon Working Papers*, 1-45.

Baayen, R.Harald, Hans Van Halteren, Anneke Neijt and Fiona Tweedie
2002 An experiment in authorship attribution. In: *Proceedings of JADT 2002.* St. Malo: Université de Rennes, 29-37.

Baayen, R. Harald, Hans van Halteren and Fiona Tweedie
1996 Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing* 11, 121-131.

Bauer, Laurie
2001 *Morphological Productivity.* Cambridge: Cambridge University Press.

Biber, Douglas
1988 *Variation across Speech and Writing.* Cambridge: Cambridge University Press.

Biber, Douglas
1995 *Dimensions of Register Variation.* Cambridge: Cambridge University Press.

Browman, Catherine and Louis Goldstein
1995 Dynamics and articulatory phonology. In Van Gelder, Tim and Robert Port (eds.), *Mind as Motion* Cambridge, Massachusetts: MIT Press, 175-193.

Burrows, John F.
1986 Modal verbs and moral principles: An aspect of Jane Austen's style.

Literary and Linguistic Computing 1, 9-23.

Burrows, John F.
1987 Word-patterns and story-shapes: The statistical analysis of narrative style. *Literary and Linguistic Computing* 2, 61-70.

Burrows, John F.
1992a Computers and the study of literature. In Butler, Christopher S. (ed.), *Computers and Written Texts.* Oxford: Blackwell, 167-204.

Burrows, John F.
1992b Not unless you ask nicely: The interpretative nexus between analysis and information. *Literary and Linguistic Computing* 7, 91-109.

Burrows, John F.
1993a Noisy signals? or signals in the noise? In: *ACH-ALLC Conference Abstracts.* Georgetown, 21-23.

Burrows, John F.
1993b Tiptoeing into the infinite: testing for evidence of national differences in the language of English narrative. In Hockey, Susan and Ide, Nancy (eds.), *Research in Humanities Computing '92.* London: Oxford University Press.

Bybee, Joan L.
2001 *Phonology and Language Use.* Cambridge: Cambridge University Press.

Bybee, Joan L.
2005 Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. manuscript submitted for publication.

Cambier-Langeveld, Tina
2000 *Temporal Marking of Accents and Boundaries.* Amsterdam: LOT.

Chatterjee, Samprit, Ali S. Hadi and Bertram Price
2000 *Regression Analysis by Example.* New York: John Wiley & Sons.

Church, Kenneth W. and Patrick Hanks
1990 Word association norms, mutual information, and lexicography. *Computational Linguistics* 16, 22-29.

Ernestus, Mirjam
2000 *Voice assimilation and segment reduction in casual Dutch: A Corpus-Based Study of the Phonology-Phonetics Interface.* Utrecht: LOT.

Ernestus, Mirjam, R. Harald Baayen, Robert Schreuder

2002 The recognition of reduced word forms. *Brain and Language* 81, 162-173.

Fougeron, Cécile and Patricia Keating
1997 Articulatory strengthening at the edges of prosodic domains. *Journal of the Acoustical Society of America* 101 (6), 3728-3740.

Gordon, Elizabeth, Margaret Maclagan and Jennifer Hay
forthcoming The ONZE Corpus. In Beal, Joan, Karen Corrigan and Hermann Moisl (eds.), *Models and Methods in Handling of Unconventional Digital Corpora.* Vol. 2. Palgrave.

Gregory, Michelle, William D. Raymond, Alan Bell, Eric Fosler-Lussier and Daniel Jurafsky
1999 The effects of collocational strength and contextual predictability in lexical production. *Chicago Linguistic Society* 35, 151-166.

Grondelaers, Stefan, Katrien Deygers, Hilde van Aken, Vicky van den Heede and Dirk Speelman
2000 Het CONDIV-corpus geschreven Nederlands. *Nederlandse Taalkunde* 5 (4), 356-363.

Harrell, Frank E.
2001 *Regression Modeling Strategies.* Berlin: Springer.

Hay, Jennifer and Andrea Sudbury
to appear How rhoticity became /r/-sandi. *Language.*

Heine, Bernd and Tania Kuteva
2005 *Language Contact and Grammatical Change.* Cambridge: Cambridge University Press.

Holmes, David. I.
1994 Authorship attribution. *Computers and the Humanities* 28 (2), 87-106.

Johnson, Keith
2004 Massive reduction in conversational American English. In: *Spontaneous Speech: Data and Analysis. Proceedings of the 1st Session of the 10th International Symposium.* Tokyo, Japan: The National International Institute for Japanese Language, 29-54.

Jurafsky, Daniel, Alan Bell, Michelle Gregory and William D. Raymond
2001 Probabilistic relations between words: Evidence from reduction in lexical production. In Bybee Joan and Paul Hopper (eds.), *Frequency and the Emergence of Linguistic Structure.* Amsterdam/Philadelphia: John Benjamins, 229-254.

Kemps, Rachèl, Mirjam Ernestus, Robert Schreuder and R. Harald Baayen
2004 Processing reduced word forms: The suffix restoration effect. *Brain and Language* 90, 117-127.

Manning, Chris and Hinrich Schutze
1999 *Foundations of Statistical Natural Language Processing.* Cambridge: MIT Press.

Norris, Dennis G.
1994 Shortlist: A connectionist model of continuous speech recognition. *Cognition* 52, 189-234.

Oostdijk, Nelleke
2002 The design of the Spoken Dutch Corpus. In Peters, Pam, Peter Collins, Adam Smith (eds.), *New Frontiers of Corpus Research.* Amsterdam: Rodopi, 105-112.

Pinheiro, José C. and Douglas M. Bates
2000 *Mixed-Effects Models in S and S-PLUS.* Statistics and Computing. New York: Springer.

Plag, Ingo, Christiane Dalton-Puffer and Baayen, R. Harald
1999 Productivity and register. *Journal of English Language and Linguistics* 3, 209-228.

Pluymaekers, Mark, Mirjam Ernestus and R. Harald Baayen
submitted Lexical frequency and acoustic reduction in spoken Dutch.

Quené, Hugo, Huub van den Bergh
2004 On multi-level modeling of data from repeated measures designs: a tutorial. *Speech Communication* 43, 103-121.

Schreier, Daniel, Elizabeth Gordon, Jennifer Hay and Margaret Maclagan
2003 The regional and linguistic dimension of /hw/ maintenance and loss in early 20^{th} century New Zealand English. *English World-Wide* 24 (2), 245-270.

Schultink, Henk
1961 Produktiviteit als morfologisch fenomeen. *Forum der Letteren* 2, 110–125.

Scott, D. R.
1982 Duration as a cue to the perception of a phrase boundary. *Journal of the Acoustical Society of America* 71, 996-1007.

Van de Velde, Hans, Roeland van Houtand Marinel Gerritsen
1997 Watching Dutch change: a real time study of variation and change in standard Dutch pronunciation. *Journal of Sociolinguistics* 1, 361-391.

Van Marle, Jaap
1988 Betekenis als factor bij productiviteitsverandering. *Spektator* 17, 341-359.

Verhoeven, Jo, Guy De Pauwand Hanne Kloots
2004 Speech rate in a pluricentric language: A comparison between Dutch in Belgium and the Netherlands. *Language and Speech* 47 (3), 297-308.

Zipf, George K.
1935 *The Psycho-Biology of Language.* Boston: Houghton Mifflin.

6 Appendices

Appendix A

Selected words ending in the suffix *-lijk* from the newspaper corpus (CONDIV):

mogelijk; duidelijk; natuurlijk; eigenlijk; uiteindelijk; moeilijk waarschijnlijk; namelijk; onmiddellijk; eindelijk; verantwoordelijk; aanvankelijk; gemakkelijk; onmogelijk; persoonlijk; makkelijk; degelijk; vermoedelijk; noodzakelijk; behoorlijk; gevaarlijk; tijdelijk; voornamelijk; afhankelijk; kennelijk; eerlijk; letterlijk; aanzienlijk; werkelijk; koninklijk; opmerkelijk; redelijk; respectievelijk; gezamenlijk; wettelijk; onduidelijk; hopelijk; onafhankelijk; gedeeltelijk; daadwerkelijk; wetenschappelijk; gerechtelijk; dergelijk; toegankelijk; oorspronkelijk; landelijk; gemeentelijk; maatschappelijk; aantrekkelijk; menselijk; nadrukkelijk; stedelijk; onvermijdelijk; openlijk; verschrikkelijk; heerlijk; uitzonderlijk; geleidelijk; voorwaardelijk; tamelijk; ongelooflijk; vriendelijk; dodelijk; pijnlijk; vreselijk; herhaaldelijk; plaatselijk; vrolijk; belachelijk; schriftelijk; hoofdzakelijk; gebruikelijk; uitdrukkelijk; onbegrijpelijk; gewoonlijk; begrijpelijk; inhoudelijk; onwaarschijnlijk; feitelijk; verwonderlijk.

Most common words from the newspaper corpus (CONDIV):

de; van; het; een; in; en; dat; op; is; te; voor; met; zijn; die; niet; aan; er; maar; ik; om; als; ook; hij; bij; uit; nog; door; naar; heeft; ze; was; dan; over; tot; jaar; worden; we; of; al; wordt; meer; hebben; je; zich; geen; werd; kan; dit; zo; wat; hun; na; wel; nu; moet; tegen; twee; deze; kunnen; haar; veel; had; uur; eerste; zou; zal; nieuwe; onder; moeten; daar; andere; wil; volgens; gaat; mijn; toch; mensen; waar; gaan; zegt.

Selected words ending in the suffix *-lijk* from the corpus of spoken Dutch (CGN):

eigenlijk; natuurlijk; waarschijnlijk; tuurlijk; moeilijk; uiteindelijk; redelijk; makkelijk; duidelijk; mogelijk; verschrikkelijk; eerlijk; dadelijk; gemakkelijk; belachelijk; onmiddellijk; ongelooflijk; eindelijk; tamelijk; lelijk; persoonlijk; gevaarlijk; afhankelijk; vriendelijk; vrolijk; ongelofelijk; hopelijk; voornamelijk; degelijk; oorspronkelijk; feitelijk; pijnlijk.

Most common words from the corpus of spoken Dutch (CGN):

aan; ah; al; als; ben; bij; daar; dan; da's; dat; de; denk; die; doen; d'r; dus; echt; een; eigenlijk; en; er; gaan; gaat; ge; gewoon; goed; had; hè; heb; hebben; heeft; heel; het; hij; hoe; ie; ik; in; is; ja; je; 'k; kan; keer; maar; meer; met; mij;

mmm; moet; naar; nee; niet; nog; nu; of; oh; om; ook; op; 't; te; toch; toen; uh; uhm; van; veel; voor; want; was; wat; we; weer; weet; wel; ze; zeg; zijn; zo.

Appendix B

Table 2
Values of the coefficients as visualized in Figure 2.

WORD	NETHERL.	FLANDERS	QUALITY	NATIONAL	REGIONAL
aantrekkelijk	0.14	-0.49	0.04	-0.02	-0.17
aanvankelijk	0.21	-0.30	0.01	0.02	-0.15
aanzienlijk	0.23	-0.48	0.02	0.09	-0.25
afhankelijk	0.22	-0.45	0.05	-0.12	-0.07
begrijpelijk	0.09	-0.56	0.04	-0.05	-0.15
behoorlijk	-0.07	-0.13	-0.05	0.06	-0.14
belachelijk	-0.19	-0.25	-0.05	-0.02	-0.08
daadwerkelijk	0.37	-0.70	-0.01	0.01	-0.15
degelijk	-0.11	-0.06	-0.03	-0.01	-0.09
dergelijk	0.01	-0.34	-0.04	-0.01	-0.10
dodelijk	-0.17	-0.23	-0.06	-0.01	-0.08
duidelijk	0.10	0.09	-0.05	0.04	-0.09
eerlijk	-0.01	-0.22	-0.04	-0.02	-0.08
eigenlijk	0.25	-0.14	-0.05	-0.03	-0.02
eindelijk	0.04	-0.11	-0.16	0.13	-0.09
feitelijk	0.30	-0.81	-0.02	0.03	-0.18
gebruikelijk	0.32	-0.80	0.05	-0.08	-0.13
gedeeltelijk	-0.10	-0.22	0.03	-0.12	-0.06
geleidelijk	-0.24	-0.16	0.11	-0.05	-0.21
gemakkelijk	0.06	-0.16	0.05	0.02	-0.19
gemeentelijk	-0.52	0.13	-0.13	-0.01	-0.02
gerechtelijk	-0.66	0.28	0.07	-0.08	-0.15
gevaarlijk	-0.11	-0.10	-0.03	0.01	-0.11
gewoonlijk	-0.20	-0.27	0.00	-0.03	-0.13
gezamenlijk	0.32	-0.62	0.06	-0.03	-0.17
heerlijk	0.19	-0.58	-0.04	-0.02	-0.09

Table 2 (continued)

WORD	NETHERL.	FLANDERS	QUALITY	NATIONAL	REGIONAL
herhaaldelijk	-0.17	-0.25	0.02	-0.06	-0.12
hoofdzakelijk	-0.40	-0.06	0.05	-0.03	-0.18
hopelijk	-0.55	0.14	-0.07	-0.03	-0.06
inhoudelijk	-0.03	-0.43	0.04	-0.08	-0.12
kennelijk	0.68	-0.95	0.07	-0.03	-0.18
koninklijk	-0.25	-0.03	0.17	-0.12	-0.19
landelijk	0.73	-1.14	0.00	-0.13	-0.03
letterlijk	0.04	-0.28	-0.03	-0.06	-0.05
maatschappelijk	-0.01	-0.35	0.12	-0.20	-0.07
makkelijk	-0.03	-0.14	-0.02	-0.04	-0.07
menselijk	-0.09	-0.28	0.08	-0.12	-0.12
moeilijk	0.01	0.08	-0.01	0.02	-0.11
mogelijk	0.25	0.00	-0.03	0.02	-0.08
nadrukkelijk	0.30	-0.68	-0.01	-0.05	-0.10
namelijk	0.32	-0.35	-0.13	0.12	-0.11
natuurlijk	0.24	-0.06	-0.09	0.05	-0.05
noodzakelijk	0.01	-0.19	0.05	-0.05	-0.13
onafhankelijk	0.18	-0.50	0.03	-0.02	-0.16
onbegrijpelijk	-0.10	-0.36	-0.04	0.04	-0.16
onduidelijk	0.28	-0.59	0.06	-0.06	-0.14
ongelooflijk	-0.13	-0.27	-0.10	0.02	-0.07
onmiddellijk	-0.34	0.29	-0.05	0.06	-0.12
onmogelijk	-0.07	-0.06	-0.01	-0.02	-0.10
onvermijdelijk	-0.02	-0.36	0.12	-0.15	-0.12
onwaarschijnlijk	-0.09	-0.40	0.03	0.00	-0.19
oorspronkelijk	-0.22	-0.13	-0.01	-0.10	-0.05
openlijk	0.01	-0.39	0.09	-0.07	-0.18

Table 2 (continued)

WORD	NETHERL.	FLANDERS	QUALITY	NATIONAL	REGIONAL
opmerkelijk	0.03	-0.29	-0.01	0.01	-0.14
persoonlijk	0.16	-0.31	0.00	-0.02	-0.11
pijnlijk	-0.15	-0.25	0.02	-0.01	-0.17
plaatselijk	-0.64	0.17	0.01	-0.09	-0.08
redelijk	0.38	-0.66	-0.06	0.05	-0.14
respectievelijk	0.04	-0.32	-0.02	0.06	-0.18
schriftelijk	-0.08	-0.36	-0.04	-0.06	-0.06
stedelijk	-0.53	0.12	-0.05	-0.02	-0.09
tamelijk	0.46	-0.90	0.08	-0.03	-0.21
tijdelijk	0.03	-0.25	-0.02	0.00	-0.11
toegankelijk	-0.23	-0.11	0.00	-0.08	-0.07
uitdrukkelijk	-0.47	0.00	0.02	-0.08	-0.11
uiteindelijk	0.16	-0.07	-0.13	0.08	-0.05
uitzonderlijk	-0.40	0.01	0.10	-0.12	-0.14
verantwoordelijk	0.24	-0.33	0.06	-0.08	-0.11
vermoedelijk	-0.21	0.02	-0.05	0.04	-0.13
verschrikkelijk	-0.08	-0.30	-0.04	-0.03	-0.08
verwonderlijk	-0.22	-0.29	0.01	0.03	-0.20
voornamelijk	0.23	-0.46	0.02	-0.08	-0.07
voorwaardelijk	-0.03	-0.42	-0.11	0.04	-0.09
vreselijk	0.20	-0.63	-0.03	-0.07	-0.05
vriendelijk	-0.14	-0.28	-0.03	-0.12	-0.01
vrolijk	0.10	-0.53	0.01	-0.03	-0.13
waarschijnlijk	0.24	-0.24	-0.02	-0.02	-0.07
werkelijk	0.08	-0.33	0.09	-0.07	-0.16
wetenschappelijk	-0.16	-0.16	0.06	-0.08	-0.13
wettelijk	-0.07	-0.23	0.03	-0.06	-0.12

Table 3
Values of the coefficients as visualized in Figure 3.

WORD	NL QUALITY	ADJ. FLANDERS	ADJ. NATIONAL	ADJ. REGIONAL
aan	0.65	-0.11	0.02	0.01
al	-0.21	-0.10	0.08	0.07
alleen	-0.98	-0.33	-0.13	-0.06
als	0.66	-0.34	-0.22	-0.12
andere	-0.87	-0.23	-0.10	-0.04
bij	0.34	-0.21	0.01	0.01
daar	-0.99	-0.20	0.07	0.10
dan	0.08	-0.17	-0.04	-0.01
dat	1.47	-0.23	-0.14	-0.05
de	3.11	-0.14	-0.02	-0.03
deze	-0.58	-0.26	-0.07	-0.04
die	1.14	-0.22	-0.20	-0.11
dit	-0.43	-0.19	0.04	0.02
door	0.38	-0.35	-0.15	-0.11
drie	-1.28	-0.09	0.15	0.13
een	2.14	-0.15	-0.08	-0.07
eerste	-1.00	-0.02	0.19	0.10
en	1.90	-0.10	-0.08	-0.09
er	0.46	-0.08	0.07	0.06
gaan	-1.14	-0.24	0.05	0.08
gaat	-0.99	-0.27	-0.03	0.03
geen	-0.28	-0.16	-0.02	0.02
haar	-0.65	-0.21	-0.08	-0.02
had	-0.71	-0.24	-0.05	-0.02
hebben	-0.18	-0.25	-0.02	0.04
heeft	0.31	-0.37	-0.13	-0.02

Table 3 (continued)

WORD	NL QUALITY	ADJ. FLANDERS	ADJ. NATIONAL	ADJ. REGIONAL
het	2.26	-0.18	-0.08	-0.03
hij	0.55	-0.26	-0.14	-0.06
hun	-0.45	-0.15	-0.06	-0.03
ik	0.32	0.03	0.11	0.09
in	2.03	-0.17	-0.05	-0.04
is	1.40	-0.27	-0.12	-0.04
je	-0.22	-0.26	-0.07	0.05
kan	-0.46	-0.11	0.02	0.03
kunnen	-0.67	-0.16	0.04	0.04
maar	0.57	-0.16	-0.05	-0.03
meer	-0.17	-0.18	-0.03	0.01
met	1.11	-0.16	0.00	-0.02
mijn	-1.30	0.00	0.26	0.16
moet	-0.57	-0.17	0.02	0.05
moeten	-0.98	-0.16	0.06	0.07
na	-0.58	-0.11	0.12	0.05
naar	0.13	-0.15	0.05	0.06
niet	0.92	-0.17	-0.07	-0.01
nieuwe	-0.91	-0.18	0.02	0.04
nog	0.06	-0.07	0.15	0.11
nu	-0.54	-0.13	0.06	0.04
of	0.06	-0.25	-0.23	-0.14
om	0.38	-0.08	0.07	0.05
omdat	-1.10	-0.24	-0.02	0.04
onder	-0.78	-0.31	-0.11	-0.06
ook	0.43	-0.17	0.01	0.02
op	1.28	-0.10	0.06	0.01

Table 3 (continued)

WORD	NL QUALITY	ADJ. FLANDERS	ADJ. NATIONAL	ADJ. REGIONAL
over	0.26	-0.34	-0.26	-0.16
te	1.26	-0.24	-0.08	-0.06
tegen	-0.67	-0.09	0.12	0.04
toch	-1.23	-0.07	0.13	0.11
tot	-0.07	-0.19	-0.03	-0.02
twee	-0.73	-0.08	0.15	0.11
uit	0.16	-0.11	0.06	0.05
van	2.46	-0.17	-0.14	-0.13
veel	-0.64	-0.31	-0.09	-0.03
volgens	-0.90	-0.31	-0.07	0.00
voor	1.13	-0.11	0.00	-0.01
waar	-1.03	-0.29	-0.07	0.00
was	0.04	-0.14	0.04	0.02
wat	-0.42	-0.16	-0.08	-0.03
we	-0.53	0.06	0.28	0.26
wel	-0.52	-0.20	0.04	0.05
werd	-0.69	0.09	0.27	0.13
wil	-1.00	-0.18	0.05	0.09
worden	-0.02	-0.24	-0.06	-0.02
wordt	-0.17	-0.20	-0.02	0.01
zal	-0.87	-0.14	0.13	0.09
ze	0.22	-0.18	-0.04	0.02
zegt	-1.14	-0.17	0.11	0.12
zich	-0.13	-0.27	-0.14	-0.08
zijn	1.20	-0.23	-0.11	-0.07
zo	-0.33	-0.20	-0.11	-0.10
zou	-0.73	-0.24	-0.10	-0.05

Table 4
Values of the coefficients as visualized in Figure 4.

WORD	WORD FREQ.	ADJ. FLANDERS	ADJ. MEN	ADJ. EDU-NON-HIGH
afhankelijk	-0.96	0.07	0.21	-0.29
belachelijk	-0.01	-0.05	0.03	-0.11
dadelijk	0.55	-1.47	-0.38	0.34
degelijk	-1.83	0.45	0.09	-0.15
duidelijk	0.64	-0.29	0.27	-0.37
eerlijk	0.32	-0.53	0.06	-0.14
eigenlijk	3.58	0.55	0.01	-0.21
eindelijk	-0.04	-0.29	-0.20	0.13
feitelijk	-2.57	1.23	0.30	-0.37
gemakkelijk	-1.24	2.17	0.16	-0.30
gevaarlijk	-0.92	0.23	0.00	-0.07
hopelijk	-1.80	1.15	0.14	-0.23
lelijk	-0.69	-0.12	-0.26	0.20
makkelijk	1.06	-1.95	-0.12	0.08
moeilijk	1.52	0.40	-0.05	-0.09
mogelijk	0.51	0.18	0.12	-0.23
natuurlijk	3.55	-0.42	0.11	-0.27
ongelofelijk	-1.55	0.51	0.40	-0.48
ongelooflijk	-0.93	1.14	0.50	-0.63
onmiddellijk	-1.75	2.32	0.54	-0.69
oorspronkelijk	-1.77	0.34	0.04	-0.10
persoonlijk	-0.93	0.47	0.09	-0.17
pijnlijk	-2.50	1.13	0.19	-0.26
redelijk	0.55	0.56	0.25	-0.38
tamelijk	-2.10	2.11	0.61	-0.74
tuurlijk	1.87	-0.03	-0.03	-0.11

Table 4 (continued)

WORD	WORD FREQ.	ADJ. FLANDERS	ADJ. MEN	ADJ. EDU-NON-HIGH
uiteindelijk	0.81	0.92	0.09	-0.23
verschrikkelijk	0.59	0.01	-0.19	0.10
voornamelijk	-1.14	-0.43	0.11	-0.16
vriendelijk	-1.51	1.22	-0.35	0.27
vrolijk	-0.87	-1.39	-0.20	0.19
waarschijnlijk	1.77	0.82	0.09	-0.26

Table 5
Values of the coefficients as visualized in Figure 5.

WORD	WORD FREQ.	ADJ. FLANDERS	ADJ. MEN	ADJ. EDU-NON-HIGH
aan	-0.0337	0.0086	0.0002	0.0012
ah	-0.1088	0.1307	-0.0010	0.0080
al	-0.0316	0.0157	-0.0059	0.0029
als	-0.0047	0.0025	0.0000	-0.0014
ben	-0.0535	-0.0138	-0.0039	0.0032
bij	-0.0261	-0.0159	-0.0036	0.0023
daar	-0.0003	0.0130	0.0016	-0.0034
dan	0.0815	-0.0078	-0.0072	-0.0001
da's	-0.0454	0.0398	-0.0008	0.0021
dat	0.1386	0.0278	0.0022	-0.0025
de	0.0756	0.0012	0.0148	-0.0056
denk	-0.0519	-0.0004	-0.0031	-0.0012
die	0.0838	-0.0099	0.0083	-0.0033
doen	-0.0526	0.0067	-0.0065	0.0011
d'r	-0.0253	-0.0241	0.0031	-0.0014
dus	0.0313	-0.0050	0.0020	-0.0103
echt	-0.0198	-0.0193	-0.0097	-0.0051
een	0.0863	-0.0184	0.0105	-0.0083
eigenlijk	-0.0670	0.0296	0.0006	-0.0076
en	0.1246	0.0079	-0.0052	0.0011
er	-0.0386	0.0095	0.0024	-0.0051
gaan	-0.0547	0.0245	-0.0042	0.0022
gaat	-0.0646	0.0096	-0.0047	0.0031
ge	-0.1657	0.1785	-0.0025	0.0150
gewoon	-0.0111	-0.0563	-0.0078	-0.0045
goed	-0.0330	0.0036	-0.0035	0.0008

Table 5 (continued)

WORD	WORD FREQ.	ADJ. FLANDERS	ADJ. MEN	ADJ. EDU-NON-HIGH
had	-0.0228	-0.0237	-0.0124	0.0032
hé	-0.0164	0.0660	0.0041	0.0044
heb	0.0122	-0.0203	-0.0032	0.0006
hebben	-0.0245	-0.0078	-0.0021	-0.0050
heeft	-0.0481	0.0062	-0.0072	-0.0021
heel	-0.0188	-0.0275	-0.0107	-0.0061
het	-0.0464	0.0694	0.0009	-0.0137
hij	-0.0317	0.0033	-0.0077	-0.0018
hoe	-0.0565	-0.0020	-0.0018	0.0029
ie	-0.0270	-0.0868	-0.0047	-0.0008
ik	0.1339	-0.0240	-0.0071	0.0019
in	0.0371	0.0056	0.0053	-0.0060
is	0.0777	0.0125	0.0032	-0.0040
ja	0.2067	0.0064	0.0027	-0.0052
je	0.1008	-0.1137	-0.0008	-0.0149
’k	0.0079	0.0422	-0.0160	0.0155
kan	-0.0313	-0.0183	-0.0061	-0.0008
keer	-0.0663	0.0166	-0.0054	0.0052
maar	0.0909	0.0012	-0.0056	0.0008
meer	-0.0550	0.0023	-0.0042	0.0007
met	0.0036	0.0035	-0.0003	0.0019
mij	-0.0630	0.0077	-0.0027	-0.0005
mmm	-0.0534	-0.0161	-0.0012	-0.0024
moet	-0.0160	0.0039	-0.0037	0.0028
naar	-0.0346	0.0122	-0.0050	0.0044
nee	0.0426	-0.0250	0.0014	-0.0001
niet	0.0747	0.0084	-0.0096	0.0040

Table 5 (continued)

WORD	WORD FREQ.	ADJ. FLANDERS	ADJ. MEN	ADJ. EDU-NON-HIGH
nog	0.0149	0.0068	-0.0035	0.0027
nu	-0.0677	0.0635	-0.0033	-0.0016
of	0.0163	-0.0080	0.0016	-0.0050
oh	0.0369	-0.1126	-0.0210	0.0043
om	-0.0430	0.0077	0.0011	-0.0069
ook	0.0656	-0.0242	-0.0089	-0.0027
op	0.0122	-0.0053	0.0036	-0.0028
't	0.0970	-0.0114	-0.0033	0.0029
te	-0.0159	0.0130	0.0024	-0.0054
toch	-0.0176	-0.0047	-0.0075	0.0016
toen	-0.0225	-0.0653	-0.0123	0.0050
uh	0.1376	-0.0638	0.0230	-0.0106
uhm	-0.0478	0.0159	0.0034	-0.0114
van	0.0418	0.0132	0.0056	-0.0015
veel	-0.0625	0.0089	-0.0028	-0.0048
voor	-0.0187	0.0141	0.0005	-0.0004
want	0.0000	-0.0189	-0.0180	0.0030
was	0.0140	0.0045	-0.0052	0.0006
wat	0.0057	-0.0175	0.0055	-0.0015
we	0.0052	-0.0091	-0.0006	-0.0035
weer	-0.0360	-0.0485	-0.0060	0.0036
weet	-0.0172	-0.0021	-0.0126	0.0070
wel	0.0584	-0.0271	-0.0071	0.0002
ze	0.0341	0.0031	-0.0165	0.0081
zeg	-0.0528	-0.0055	-0.0152	0.0109
zijn	-0.0230	0.0296	0.0031	-0.0054
zo	0.0298	0.0210	-0.0110	0.0054

Table 6
Values of the coefficients as visualized in Figure 7.

WORD	NETHERLANDS	FLANDERS
afhankelijk	-1.65	0.58
dadelijk	-13.99	10.05
duidelijk	-3.68	1.43
eerlijk	-0.70	1.93
eigenlijk	-2.18	-0.27
eindelijk	-4.42	0.57
moeilijk	0.00	0.38
mogelijk	-1.27	0.42
natuurlijk	-2.86	4.17
persoonlijk	-0.32	0.48
tuurlijk	-1.73	10.76
uiteindelijk	-13.91	9.74
vriendelijk	-1.82	-0.83
waarschijnlijk	-1.23	2.29