Social, Geographical, and Register Variation in Dutch: From Written MOGELIJK to Spoken MOK

Karen Keune *(karen.keune@mpi.nl)*
University of Nijmegen
Mirjam Ernestus *(mirjam.ernestus@mpi.nl)*
Max Planck Institute for Psycholinguistics
Roeland van Hout *(r.v.hout@let.ru.nl)*
University of Nijmegen
Harald Baayen *(baayen@mpi.nl)*
University of Nijmegen

In spontaneous speech words are often pronounced in reduced form. Some words are reduced to such an extent that an orthographic transcription would be very different from the orthographic norm. An example from Dutch is the word *MOGE-LIJK* ('poss-ible'), which can be pronounced not only as MO.GE.LEK but also as MO.GEK, MO.LEK, or even as MOK.

Strongly reduced word forms are difficult to interpret without syntactic or semantic context. When speakers of Dutch are presented with the word *mok* in isolation, they tend not to be able 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 *MOGELIJK*. 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 (Kemps et al.).

An important predictor for the degree of reduction in speech production is lexical frequency, as demonstrated by Jurafsky et al.. 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. Furthermore, the degree of reduction is modulated by the extent to which a word is predictable from its context. In addition, frequency of occurrence has been shown to affect the realization of word final dental plosives in monomorphemic words (Bybee), and a negative correlation between frequency and acoustic length has been observed for several kinds of derived words in Dutch, including words with the suffix *-LIJK* (Pluymaekers et al.).

It is an open question to what extent the use of reduced forms is co-determined by social, geographical, and stylistic factors. Various corpus-based studies have shed light on variation in the use of language. Biber 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 (Biber). In the domain of literary studies, Burrows demonstrated regional differences in English narrative, diachronic change in literary texts, and even sex-specific differences in the writing of English historians born before 1850 (see, e.g. Burrows). Studies in authorship attribution revealed, furthermore, that differences in speech habits can sometimes be traced down even to the level of individual language users (Holmes; Baayen et al.). Finally, it has been shown that derivational affixes are used to a different extent in spoken and written registers (Baayen; Plag et al.).

The aim of the present study is to investigate the extent to which the use of words in *-LIJK* varies systematically as a function of speech register, the speaker's sex, level of education, and of whether the speaker lives in Flanders or in the Netherlands. For spoken Dutch, we address the more specific question to what degree these factors (and contextual predictability) codetermine the extent to which words in *-LIJK* are reduced.

We first studied the social and geographic variation in the frequency of use of words in *-LIJK* in a corpus of Dutch newspapers. We selected all occurences of 80 high-frequency words in *-LIJK* from seven newpapers using a 2 by 3 factorial design. We distinguished between Flemish and Dutch newspapers (Country) and contrasted quality newspapers, national newspapers, and regional newspapers (Register). In parallel, we conducted a study using the same design based on the 80 most frequent function words (pronouns, auxiliaries, connectives, determiners, numerals, etc.), following Burrows. In both analyses, we observed significant and remarkably similar regional and stylistic differentiation. This suggests that the syntactic habits of journalists (as revealed by their use of function words) are consistent with their habits with respect to the use of adverbs and adjectives in *-LIJK*.

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, unscripted speech in the Corpus of Spoken Dutch (CGN), using a 2x2x2 factorial design in which we contrasted speakers from Flanders with speakers from the Netherlands (Country), men with women (Sex), and highly eductated with less educated speakers (Education). As before, we carried out a parallel study using the most frequent function words. This time, we observed a marked difference between the function 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 function words, by contrast, revealed that men made less use of function words compared to women, suggesting a slightly higher information density (carried by content words) for men. In addition to these main effects, we observed marked (and significant) differences in how individual function 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 social and regional variation in the degrees of reduction of words in *-LIJK* for 14 words that occurred sufficiently often in the different subcorpora of the CGN defined by our factorial design contrasting Country, Sex and Education, and that revealed substantial degrees of reduction. Two transcribers classified the degree of reduction for a total of 946 tokens. We considered two kinds of reduction, one primarily affecting the suffix, the other affecting the vowel in the word initial syllable. Both analyses show that in Flanders speakers reduce less than in the Netherlands. The reduction involving the suffix is more prominent for men compared to women. Moreover, highly educated Flemish speakers use fewer reduced forms than do less highly educated Flemish speakers. Finally, there were significant differences in the extent to which the individual words underwent reduction that we could trace back to the speaker's region.

In addition to these social and regional factors, the degree of reduction was significantly co-determined by two linguistic factors: the word's position in the sentence, and the extent to which the word is predictable from its context. We used the Mutual Information measure to gauge contextual predictivity. Words in *-LIJK* with a high mutual information, i.e., words that exhibited a high degree of predictability from the preceding word, revealed more reduction: As the information load of a word in *-LIJK* decreases, its formal distinctiveness in production decreases as well. In this respect, highly-reduced and semantically opaque forms in *-LIJK* such as *TUUK* (for *NATUURLIJK*, 'of course') and *EIK* (for *EIGENLIJK*, 'in fact') are becoming similar to function words. 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.

For our analyses, we made extensive use of multilevel modeling of covariance, a statistical technique that offers two advantages compared to principal components analysis, factor analysis, and correspondence analysis. First of all, multilevel modeling allows the researcher to directly assess the significance of the predictors in the model, as well as how the individual words 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.

Although derived words are generally classified as open-class words, as opposed to the closed class function words, it is noteworthy that the suffix *-LIJK* is hardly productive. Furthermore, we have shown that if the information load of a word in *-LIJK* decreases, its formal distinctiveness in production decreases as well. Thus, high-frequency forms in *-LIJK* are becoming more similar to function words with respect to their lack of productivity and compositionality, with respect to their being social and stylistic markers, and with respect to their acoustic form.

Bibliography

Baayen, R.H, H. Van Halteren, and F. Tweedie. "Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution." *Literary and Linguistic Computing* 11 (1996): 121-131.

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

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

Burrows, J.F. "Computers and the study of literature." *Computers and Written Texts.* Ed. C.S. Butler. Oxford: Blackwell, 1992. 167-204.

Bybee, J.L. *Phonology and language use.* Cambridge: Cambridge University Press, 2001.

Holmes, D.I. "Authorship attribution." *Computers and the Humanities* 28.2 (1994): 87-106.

Jurafsky, D., A. Bell, M. Gregory, and W.D. Raymond. "Probabilistic relations between words: Evidence from reduction in lexical production." *Frequency and the emergence of linguistic structure.* Ed. J.L. Bybee and P. Hopper. Amsterdam: John Benjamins, 2001. 229-254.

Kemps, R, M. Ernestus, R. Schreuder, and R.H. Baayen. "Processing reduced word forms: The suffix restoration effect." *Brain and Language* 90 (2004): 117-127.

Plag, I., C. Dalton-Puffer, and R.H. Baayen. "Productivity and register." *Journal of English Language and Linguistics* 3 (1999): 209-288.

Pluymaekers, M., M. Ernestus, and R.H. Baayen. *Lexical frequency and acoustic reduction in spoken Dutch.* In preparation.