Computational Linguistics in the Netherlands Journal 4 (2014) 29-38 Submitted 05/2014; Published 12/2014

Social, geographical, and lexical influences on Dutch dialect pronunciations

Vinnie Ko[*] V.B.KO@STUDENTS.UU.NL
Martijn Wieling[***] M.B.WIELING@RUG.NL
Ernst Wit[**] E.C.WIT@RUG.NL
John Nerbonne[***] J.NERBONNE@RUG.NL
Wim Krijnen[**] W.P.KRIJNEN@RUG.NL

[*] *Department of Methodology and Statistics, Utrecht University*

[**] *Department of Statistics and Probability, University of Groningen*

[***] *Department of Information Science, University of Groningen*

Abstract

Wieling et al. (2011) combined generalized additive modeling (GAM) with mixed-effects regression modeling to identify the influence of social, lexical, and geographical variables on the variation of Dutch dialect pronunciations. The conclusion of their study was that the pronunciation distance from standard Dutch became greater for locations with a smaller population, a higher average age (of the inhabitants), words with a greater frequency, and words with relatively many vowels. When Wieling et al. (2011) performed their quantitative study in 2011, they were not able to analyze the dataset in a single generalized additive mixed-effects regression model due to the large size of the dataset. Instead, they first used a generalized additive model to represent geography and included the fitted values of this non-linear model as a predictor in a linear mixed-effects regression model. As more advanced methods to fit generalized additive mixed-effects regression models have become available, we improve on their approach here by constructing a single generalized additive (i.e. non-linear) mixed-effects regression model in which the non-linear geographical influence is varied depending on word frequency and word category (i.e. verbs vs. non-verbs). Non-verbs and higher frequency words generally showed a higher pronunciation distance from standard Dutch than lower frequency words and verbs. In contrast to Wieling et al. (2011), we did not find enough support to include the number of inhabitants and the average income in a location in our model. However, we did find a comparable effect of the vowel-consonant ratio. Our findings highlight the potential of using generalized additive modeling to uncover significant non-linear patterns, while simultaneously allowing for the inclusion of regular (linear) predictors and an extensive random-effects structure.

1. Introduction

Studies of language variation face two fundamental problems in data analysis. The first problem concerns which linguistic variables to analyze among a potentially wide range of variables. There is a sociolinguistics tradition (Bayley 2002) which selects a small number of categorical linguistic variables and uses logistic regression to test hypotheses about whether they are associated with extralinguistic factors such as age, sex or social class. The problem with this approach is the somewhat arbitrary selection of variables, given how complex languages are: involving several dozen basis sounds (phonemes), several tens of thousands of words, and several hundred morphological and syntactic rules or constructions (Nerbonne 2009). Dialectometry arose *inter alia* as an answer to this problem of "cherry picking" (Goebl 1984) by focusing on the analysis of large aggregates of (mostly categorical) data, but many practitioners of dialectology were dissatisfied with the emphasis on large aggregates and criticized dialectometry for its lack of attention to linguistic detail (Schneider 1988). Because mixed-effects regression models allow one to treat an aggregate measure as the dependent

variable and simultaneously examine the contributions of the individual elements via the random-effects structure, they offer at least a step towards a solution to this problem.

A second problem concerns the treatment of the explanatory variable geography (see Wieling and Nerbonne 2015), where the most popular approach has been to operationalize geography as a simple distance (Nerbonne and Heeringa 2007), enabling the testing of hypotheses concerning how much influence geography has on language variation (Nerbonne 2013). It is clear that information is lost when projecting two dimensions (or even three if one considers elevation) to a single one, and this has led to approaches in which two dimensions are considered (Grieve et al. 2011). These approaches are clearly superior to distance-based models with respect to their treatment of geography, but only a limited number of studies (Wieling et al. 2011, Wieling et al. 2014) have tried to use an approach integrating both geography, social variables (e.g., age, social class) and linguistic variables (e.g., syntactic class of words, vowel-consonant ratio). Generalized additive models (GAMs) are capable of analyzing the two dimensions of geographic influence (and their non-linear interaction) and — in combination with the mixed-effects regression approach — are able to incorporate independent variables to gauge their influence on pronunciation variation in an integrated way.

The present paper aims to further the research line in which generalized additive mixed-effects regression models are used to analyze linguistic variation.

2. Materials and methods

2.1 Pronunciation and social data

The pronunciation distances and social data from the study of Wieling et al. (2011) were re-used in this study and obtained from the paper package associated with that paper stored at the Potsdam Mind Research Repository (`http://openscience.uni-leipzig.de`). The dataset contains the pronunciation distances from standard Dutch for 559 words for 424 locations in the Netherlands. The pronunciation distances were obtained using the PMI-based Levenshtein distance (Wieling et al. 2012) applied to the phonetic transcriptions from the Goeman-Taeldeman-Van Reenen-Project (Taeldeman and Goeman 1996). For example, as reported by Wieling et al. (2011), the pronunciation distance between two dialectal variants of the Dutch word 'binden' [bɪndən] and [bɛində] is:

bɪndən	insert ɛ	0.034
bɛmdən	subst. i/ɪ	0.020
bɛindən	delete n	0.024
bɛində		
		0.078

Besides the pronunciation distances, the dataset includes information about the locations (i.e. longitude, latitude, number of inhabitants, average income and average age of the inhabitants), speakers (age and gender), and words (i.e. word category, word frequency and vowel-consonant ratio). More details about this dataset can be found in Wieling et al. (2011).

2.2 Random-effect factors

Wieling et al. (2011) used three random-effect factors in their analysis: word, location (essentially overlapping with speaker, as only a single speaker was recorded per location) and transcriber. How different a word's pronunciation is from standard Dutch varies markedly across words, and this can be captured by including a random intercept per word. For instance, the word *donder* 'thunder' is pronounced in very different ways at different locations, and generally has a relatively high pronunciation distance from standard Dutch. In contrast, the word *bitter* 'bitter' shows very little dialect variation, and generally has a relatively low pronunciation distance from standard Dutch. Even though location is essentially already included in the fixed-effects part of the model as a combination of longitude and latitude, we assessed if it was necessary to include it as a random-effect

factor as well (i.e. capturing the differences between nearby locations). Besides random intercepts, which model the structural variability in the pronunciation distances from standard Dutch per word, location or transcriber, we assessed the presence of random slopes for the predictors included in our model. These random slopes allow the effect of the predictor to vary for each level of the random-effect factor. For example, the effect of the average age of the inhabitants in a location on the pronunciation distance from standard Dutch might vary per word. This approach is in line with Wieling et al. (2011) and prevents overconfident p-values.

2.3 Generalized additive mixed-effects modeling

Wieling et al. (2011) initially tried to apply generalized additive mixed-effects regression modeling, which is a non-linear model that analyzes both fixed effects and random effects. An important benefit of a generalized additive model is that the type of non-linearity does not have to be specified in advance, but is determined automatically during model fitting in a way that prevents overfitting (via cross-validation). Unfortunately, in 2011 creating a large generalized additive mixed-effects regression model was too computationally expensive, which led Wieling and colleagues to apply a two-step analysis consisting of fitting a simple generalized additive model to represent geography and using the fitted values of this model as predictor in a linear mixed-effects regression model. The main disadvantage of this two-step analysis is that it cannot describe complex non-linear relationships (i.e. interactions between geography and other predictors).

Fortunately, in 2013 a new and very efficient function was developed to induce generalized additive mixed-effects regression models: `bam`, available in the `R` package `mgcv` (Wood et al. 2014). By using this function, we were able to include both fixed and random effects in one generalized additive model, and this is comparable to the approach followed by Wieling et al. (2014) in their study investigating Tuscan lexical variation.

2.4 Measures of model fit

For the estimation of the non-linear and random effect terms there are three performance criteria available in the `bam` approach: `GCV` (generalized cross validation), `ML` (maximum likelihood estimation) and `fREML` (fast restricted maximum likelihood estimation). As `bam` reports negative likelihoods, a model with a lower score fits the data better (i.e. the differences between the fitted values and the observed values are small) and generalizes to a greater extent (i.e. it performs well on new data) than a model with a higher score. Although all three criteria share this same idea of model performance, there are some technical differences among them. `GCV` is an approximation of the out-of-set error obtained by scaling the original residuals by an extrapolation factor, `ML` uses the maximum likelihood principle to judge the model performance, and `fREML` measures the fit of the variance parameters and uses the scaled average of the likelihood over all possible values of the estimated coefficients. Models which differ in their fixed effects can only be compared using `ML` or `GCV`. For comparing random effects `fREML` can be used as well. As `GCV` tends to be prone to overfitting (Wood 2011), we use `ML` when comparing models differing in the fixed effects, and `fREML` for our final model and when comparing models differing in the random effects.

2.5 Model selection

One of the advantages of our single-step generalized additive mixed-effects regression analysis is that we can more easily obtain information about the overall model quality. Starting from the variables included in the model of Wieling et al. (2011), we assessed if variables needed to be omitted or added in a stepwise manner. The comparison of the models was based on the Akaike Information Criterion (AIC; Akaike 1974). The AIC score is calculated by combining the `ML` (or `fREML` score, depending on whether fixed effects or random effects are compared) and offsetting this against the complexity of the model in terms of the parameter count. Lower AIC values indicate an improved model with

a better (relative) goodness-of-fit. We omitted variables from the original model of Wieling et al. (2012) if the omission resulted in a decreased AIC value (i.e. a better fit), and we included additional variables when the inclusion resulted in a decrease of the AIC of at least 2 (i.e. to be conservative in adding new predictors). For each fixed-effect predictor included in the model, we assessed if random slopes were necessary (i.e. when the addition of the random slope resulted in an AIC decrease).

3. Results

In line with Wieling et al. (2011), we included random intercepts for the random-effect factors word, location and transcriber (the inclusion of each of these was supported by a reduction of the AIC score compared to the model without the random intercept).

In the exploratory phase (and in line with the post-hoc analysis reported by Wieling et al. 2011) we noted that verbs showed a markedly different geographical pattern compared to other word categories, and that word frequency significantly influenced this pattern (more on this below). Therefore, we allowed the effect of geography to vary depending on word category and word frequency. Once the geographical part of the model was set, we assessed the significance of the variables considered by Wieling et al. (2011). Our final model is specified as follows:

```
PronDistStdDutch.log.c ~ te(Longitude, Latitude, WordFreq.log, d = c(2, 1),
    by = WordCategory) +
    WordCategory + LocAvgAge + WordVCratio.log +
    s(Word, bs = "re") + s(Location, bs = "re") + s(Transcriber, bs = "re") +
    s(Word, LocAvgAge, bs = "re") +
    s(Location, WordVCratio.log, bs = "re") +
    s(Transcriber, WordVCratio.log, bs = "re")
```

The dependent variable in our final model is the pronunciation distance from standard Dutch (log-transformed and centered). The fixed-effects part of the model contains the following predictors: a non-linear interaction (included in a tensor product `te()`) between longitude, latitude (i.e. geography), word frequency (log-transformed, as its distribution is skewed) and word category (verbs vs. non-verbs), the average age of the inhabitants in the location where the dialect is spoken, and the ratio between vowel and consonants in a word (log-transformed). The `d`-tuple in the tensor product indicates that both longitude and latitude (the first two continuous predictors in the tensor) are on the same scale (i.e. degrees), whereas word frequency (the final continuous predictor in the tensor) is on a different scale (i.e. log-transformed counts). Since the tensor products are centered, the separate word category variable is necessary to capture a potential constant difference between the two three-dimensional surfaces. The random-effects structure is indicated by the `s(..., bs = "re")` terms. Random intercepts can be identified by an `re`-term including only a single factorial predictor. For example, `s(Word, bs = "re")` represents a random intercept per word. Similarly, random slopes contain two predictors. For example, `s(Word, PopAvgAge, bs = "re")` represents the by-word random slope for the average age per location.

All included random intercepts and slopes contributed significantly to the model. In Figure 1, visualizing the random intercepts, we can observe that the structural variability associated with words is highest. The final model explains 46.1% of the variance of the dependent variable and the results are summarized in Table 1.

3.1 Interaction between geography and lexical predictors

Wieling et al. (2011) included the effect of geography as a simple non-linear interaction between longitude and latitude. In a post-hoc analysis, they showed that the geographical pattern varied based on word category and word frequency. We first tested if there was a difference in the geographical pattern between verbs and non-verbs, and this appeared to be the case as the AIC of the model

	Estimate	t-value	edf	F-value	p-value
Intercept	-0.036	-0.791			0.429
Verb vs. non-verb	-0.030	-2.191			0.028
Average age per location	0.003	2.031			0.042
Vowel-consonant ratio (log)	0.139	8.070			< 0.001
Geographical distribution of non-verbs			87.2	16.5	< 0.001
Geographical distribution of verbs			99.3	26.7	< 0.001
Word (random intercept)			432.9	4978.2	< 0.001
Location (random intercept)			376.9	130.0	< 0.001
Transcriber (random intercept)			19.2	4085.1	< 0.001
By-word random slope for average age			196.9	2211.9	< 0.001
By-location random slope for v.c.r.			352.8	66.6	< 0.001
By-transcriber random slope for v.c.r.			22.8	3182.7	< 0.001

Table 1: Summary of the final model. Geographical distribution is represented as a tensor of longitude, latitude and word frequency (log-transformed). The vowel-consonant ratio is abbreviated by 'v.c.r.'. The log-transformation of a variable is denoted by '(log)'. The column 'edf' indicates the estimated degrees of freedom for each smooth function.

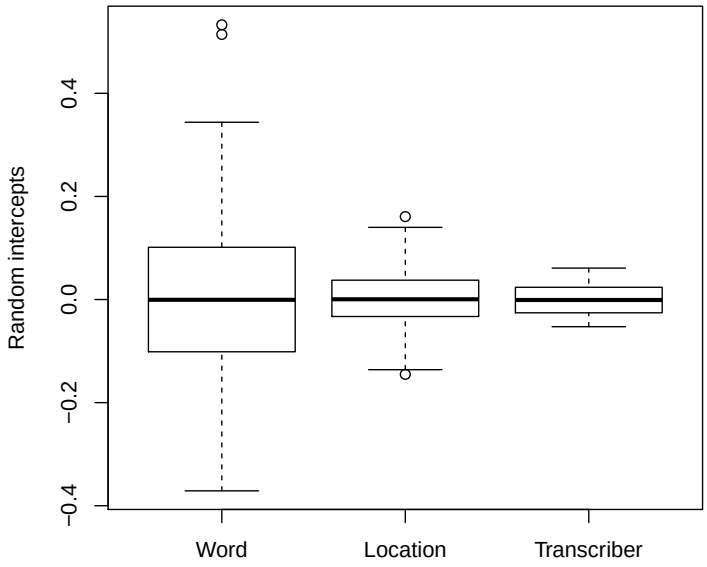

Figure 1: Box plot showing the variation associated with each of the three random-effect factors used in the model

including this contrast was 7431 units lower than the model without the contrast. Subsequently, we tested the interaction of this geographical pattern with word frequency, which again improved

the fit of the model significantly (an AIC decrease of 1918 units). The resulting interaction (as a contour plot) is shown in Figure 2.

Clearly, the central area of the Netherlands (i.e. Holland and Brabant) shows the smallest pronunciation distance from standard Dutch for both verbs and non-verbs. This confirms the general belief that standard Dutch is mostly used in the Dutch capital and its surrounding area. The peripheral areas, i.e. the northeast, southeast and southwest corners of the country, generally show the largest pronunciation distance from standard Dutch for both types of words. While Frisian is a separate language from Dutch, its pronunciation distances from standard Dutch are not always the largest. For example, for the low-frequency verbs, the Groningen region shows greater pronunciation distances.

The contrast between verbs and the non-verbs in the summary in Table 1 shows that verbs generally show a smaller pronunciation distance from standard Dutch than non-verbs. This finding is in line with the results reported by Wieling et al. (2011). However, in Figure 2 we clearly see that the variability in pronunciation distances is much higher for the verbs than for the non-verbs.

Similar to Wieling et al. (2011) we can see that the graphs associated with a high (i.e. maximum) word frequency show, in general, a lighter shade of gray than the graphs associated with the low (i.e. minimum) word frequency. This corresponds well with the reported effect of word frequency by Wieling et al. (2011), with words of higher frequency being more resistant to change to the standard language (Pagel et al. 2007). Interestingly, for higher-frequency non-verbs the pronunciations in Drenthe (in the northeastern part of the Netherlands) tend to become closer to standard Dutch.

3.2 Additional demographic and lexical predictors

As a model without population size (i.e. the number of inhabitants) per location resulted in a better model fit, we excluded this predictor from our model. While Wieling et al. (2011) reported this variable to be significant, we did not find enough support for its inclusion, likely due to the more complex influence of geography included in our model. Similarly, the average income per location was not included in our final model (but note that Wieling et al. 2011 reported this variable to be non-significant as well).

In line with the results of Wieling et al. (2011), we found that the average age in a location was a significant predictor. For locations with a higher average age, pronunciation distance tends to be higher. Another significant predictor was the vowel-consonant ratio. In general, a greater ratio predicts a greater pronunciation distance from standard Dutch. This confirms the finding of Wieling et al. (2011) and linguistically makes sense as vowels are more variable than consonants (Keating et al. 1994). Similar to Wieling et al. (2011), we did not identify a significant effect of any of the speaker-related predictors (i.e. age and gender).

4. Conclusion and discussion

In this re-analysis of the study by Wieling et al. (2011), we used a single generalized additive mixed-effects regression model. In general, we confirmed most of the findings obtained by their explanatory quantitative model incorporating geographical, social and lexical variables as independent variables. However, we uncovered some additional findings as well. With respect to geography, we showed that a non-linear interaction between word frequency (in general, higher frequency words were associated with greater pronunciation distances from standard Dutch), word category (contrasting verbs from non-verbs) and longitude and latitude resulted in the best fitting model. Our results contrasted with those of Wieling et al. (2011) in that we did not find support for the inclusion of the number of inhabitants in a location, nor the average age of the inhabitants in a location.

In this study, we used a Levenshtein-based measure as an indicator of pronunciation difference from standard Dutch. Although the Levenshtein distance is a useful tool to convert differences in pronunciation to a single number, the details of the differences are lost. For example, the dialects in

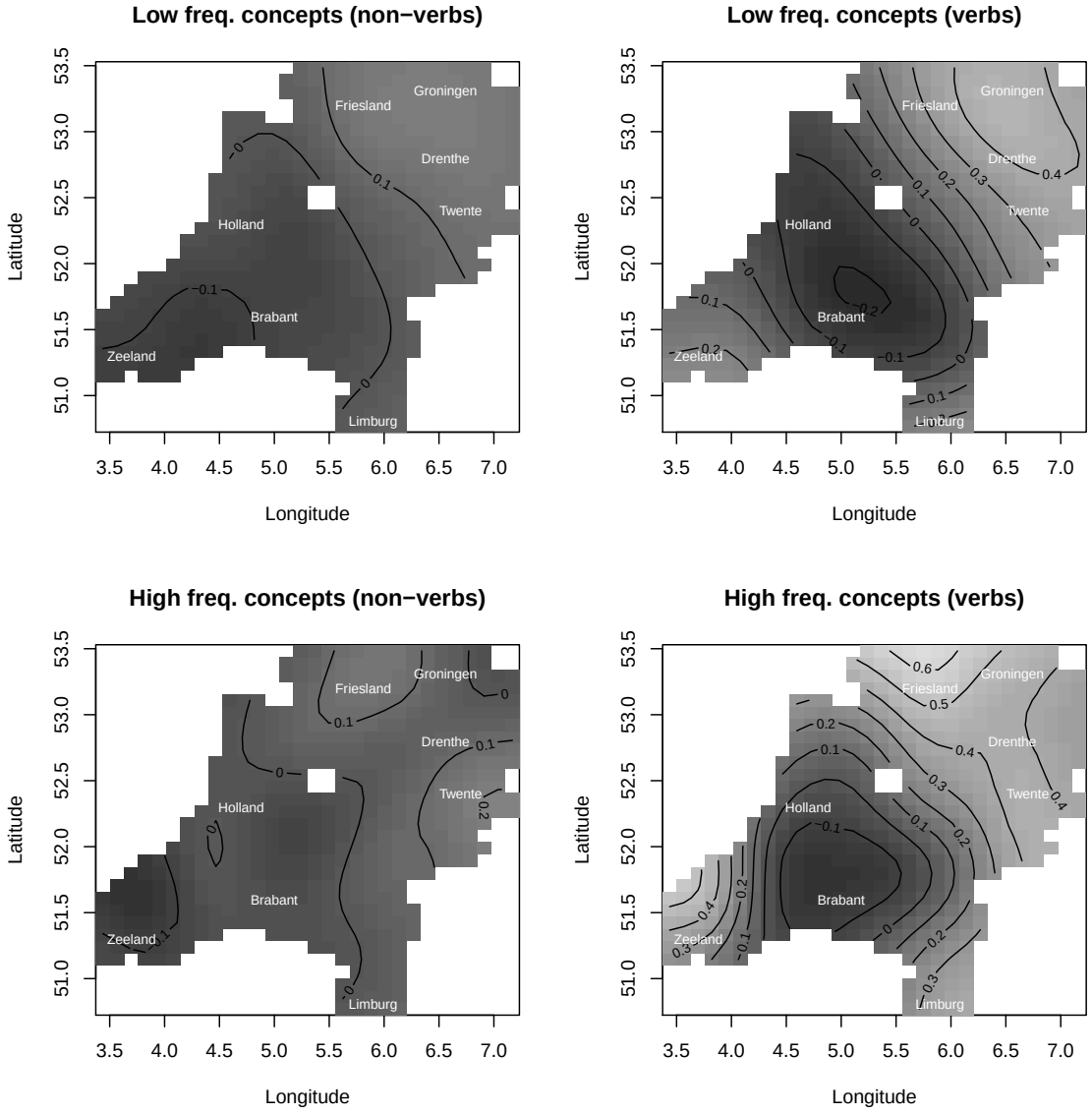

Figure 2: Contour plot of pronunciation distance (log-transformed and centered) as a function of longitude, latitude, word frequency and whether the category of the word is a verb or not. The darker the color, the smaller the pronunciation distance. The contour lines represent distance isoglosses. The empty square indicates the IJsselmeer, a large lake in the Netherlands. The title of each map shows information about the words used in the map.

the south and the north all show large pronunciation distances from standard Dutch. However, as the respective pronunciations are very different, a method in which these differences could be taken into account (perhaps by focusing on differences per segment rather than per word) would be even more insightful.

More generally, this study (in line with Wieling et al. 2014) confirms the use of the generalized additive mixed-effects regression approach as a suitable integrated method for analyzing linguistic data with an eye to simultaneously understanding geographic, social and linguistic influences on the distribution of the variation.

References

Akaike, Hirotugu (1974), A new look at the statistical model identification, *Automatic Control, IEEE Transactions on* **19** (6), pp. 716–723.

Bayley, Robert (2002), The quantitative paradigm, *in* Chambers, Jack, Peter Trudgill, and Natalie Schilling-Estes, editors, *The Handbook of Language Variation and Change*, Wiley-Blackwell, pp. 117–141.

Goebl, Hans (1984), *Dialektometrische Studien: Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF. 3 Vol*, Max Niemeyer, Tübingen.

Grieve, Jack, Dirk Speelman, and Dirk Geeraerts (2011), A statistical method for the identification and aggregation of regional linguistic variation, *Language Variation and Change* **23** (2), pp. 193–221.

Keating, Patricia A, Björn Lindblom, James Lubker, and Jody Kreiman (1994), Variability in jaw height for segments in English and Swedish VCVs, *Journal of Phonetics* **22** (4), pp. 407–422.

Nerbonne, John (2009), Data-driven dialectology, *Language and Linguistics Compass* **3** (1), pp. 175–198.

Nerbonne, John (2013), How much does geography influence language variation?, *in* Auer, Peter, Martin Hilpert, Anja Stukenbrock, and Benedikt Szmrecsanyi, editors, *Space in Language and Linguistics. Geographical, Interactional, and Cognitive Perspectives*, De Gruyter, Berlin, pp. 220–236.

Nerbonne, John and Wilbert Heeringa (2007), Geographic distributions of linguistic variation reflect dynamics of differentiation, *in* Featherston, Sam and Wolfgang Sternefeld, editors, *Roots: Linguistics in search of its evidential base*, Walter de Gruyter, Boston and Berlin, pp. 267–297.

Pagel, Mark, Quentin Atkinson, and Andrew Meade (2007), Frequency of word-use predicts rates of lexical evolution throughout indo-european history, *Nature* **449** (7163), pp. 717–720.

Schneider, Edgar (1988), Qualitative vs. quantitative methods of area delimitation in dialectology: A comparison based on lexical data from Georgia and Alabama, *Journal of English Linguistics* **21** (2), pp. 175–212.

Taeldeman, Johan and Ton Goeman (1996), Fonologie en morfologie van de Nederlandse dialecten: Een nieuwe materiaalverzameling en twee nieuwe atlasprojecten, *Taal en Tongval* **48**, pp. 38–59.

Wieling, Martijn and John Nerbonne (2015), Advances in dialectometry, *Annual Review of Linguistics*.

Wieling, Martijn, Eliza Margaretha, and John Nerbonne (2012), Inducing a measure of phonetic similarity from pronunciation variation, *Journal of Phonetics* **40** (2), pp. 307–314.

Wieling, Martijn, John Nerbonne, and R. Harald Baayen (2011), Quantitative social dialectology: Explaining linguistic variation geographically and socially, *PLOS ONE* **6** (9), pp. e23613.

Wieling, Martijn, Simonetta Montemagni, John Nerbonne, and R Harald Baayen (2014), Lexical differences between Tuscan dialects and standard Italian: A sociolinguistic analysis using generalized additive mixed modeling, *Language*.

Wood, Simon N, Yannig Goude, and Simon Shaw (2014), Generalized additive models for large data sets, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*.

Wood, S.N. (2011), Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models, *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **73** (1), pp. 3–36.