Automatically measuring the strength of foreign accents in English

Martijn Wieling[a], Jelke Bloem[b], Kaitlin Mignella[b], Mona Timmermeister[b] and John Nerbonne[b]

[a]Department of Quantitative Linguistics, University of Tübingen, [b]Department of Humanities Computing, University of Groningen

{wieling, jelke.bloem, kmignella}@gmail.com, mona2705@hotmail.com, j.nerbonne@rug.nl

Abstract

We measure the differences between the pronunciations of native and non-native American English speakers using a modified version of the Levenshtein (or string edit) distance applied to phonetic transcriptions. Although this measure is well understood theoretically and variants of it have been used successfully to study dialect pronunciations, the comprehensibility of related varieties, and the atypicalness of the speech of the bearers of cochlear implants, it has not been applied to study foreign accents. We briefly present an appropriate version of the Levenshtein distance in this paper and apply it to compare the pronunciation of non-native English speakers to native American English speech. We show that the computational measurements correlate strongly with the average "native-like" judgments given by more than 1000 native U.S. English raters (r = -0.8, $p < 0.001$). This means that the Levenshtein distance is qualified to function as a measurement of "native-likeness" in studies of foreign accent.

Key words: Foreign accent, Levenshtein distance, Edit distance, Pronunciation, Validation

1. Introduction

Most speakers of a foreign language speak with an accent, particularly if they have learned the language after childhood. Foreign accents have attracted a good deal of attention from specialists in second-language (hence: L2) learning, but also from researchers investigating whether there is a critical period within which native-like language acquisition must occur. Piske, MacKay and Flege, (2001) review a large body of literature noting that the age at which one begins learning, the time spent using the language (residence), and its amount of use may be shown to affect how native-like an accent ultimately becomes. Investigations seeking to explain the strength of foreign accents may be motivated practically, i.e. with an aim to influence second-language learning methods, but also theoretically, i.e., with an aim to understand how language is learned.

Many studies investigating foreign accents and the possible presence of a critical period in L2 learning focus on a single second language (mainly English) and only one or at most a few L1 backgrounds. This is not surprising, as obtaining foreign accented speech of people with various L1 backgrounds and the native judgments about their speech is a labor-intensive procedure. However, language background is an important determinant of foreign accent and, consequently, comparing a large set of backgrounds would be beneficial for our understanding of foreign accents (Piske, MacKay and Flege, 2001).

To facilitate this research, we propose in this paper to use the Levenshtein distance as an automatic computational method to determine how different accented speech is

from native speech. If speech samples for a few dozen words are available in a broad phonetic transcription, then the application of the Levenshtein procedure yields a numerical measure of how different the foreign-accented pronunciations are from the native pronunciations. This procedure has proven effective in measuring dialect pronunciation differences and in measuring the comprehensibility of related varieties. The Levenshtein procedure is more readily replicable than other methods of assaying pronunciation differences, and, in particular, relies less on the (subjective) selection of a small number of features whose differences are tallied. Naturally, a measurement technique must be validated before its results may be relied on, which is why we compare the results of our measurements to human judgments of accent strength in this paper.

2. Material

2.1. The Speech Accent Archive

In this study, we use data from the Speech Accent Archive (Weinberger and Kunath, 2011). The Speech Accent archive is digitally available at http://accent.gmu.edu and contains a large sample of speech samples in English from people with various language backgrounds. Each speaker reads the same paragraph of 69 words in English:

> *Please call Stella. Ask her to bring these things with her from the store: six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station.*

Below we provide (the first lines of) the transcriptions of a (i) German woman who lived for twenty-five years in the U.S. (German1 on the website), (ii) a French woman who lived in the U.S. for only two months (French3); (iii) an Italian man who lived in

the U.S. for 3-4 months (Italian2); and a Chinese woman who lived in the U.S. for 1 year (Cantonese1).

German: [pʰliːs kʰal stela æsk ɚ tu bʁĩŋ d̪iːs sĩŋks wɪθ hɜ fʁɑm d̪ə stoə siks]

French: [pʰliːz kɔl stɛle æsk ɛɹ tŭ bɹĩŋ d̪iz sɻ̩ŋs wɪð ɛɹ fɹʌm ðə stɔə sɪks]

Italian: [pliz kʰɔːl stɛlːa æsk hɜɹ tu bɹĩŋk d̪izə̆ t̪ĩŋz̩ wɛd hɜɹ fɹɔ̃m də stɔɹi sɪks]

Cantonese: [pʰɾis kʰɑl stɛlʌ as hɜɹ tŭ bɹĩŋ dis fɻ̩ɪŋs wɪf hɜɹ fɹõm də stoɹ sɪs]

We provide these examples to illustrate that the accents database contains a wealth of interesting data. The stereotypical elements of accents are present: every speaker has trouble with the interdental fricatives, but the substitutions are different (compare the Italian's pronunciation of 'things' to the others; the German devoices final obstruents, e.g., 'please'; the French speaker drops initial /h/ in 'her'; the Italian speaker adds a vowel to 'these' to create a second CV syllable; and the Cantonese speaker simplifies consonant clusters in words such as 'ask'). Also note that other stereotypical accent modifications and substitutions are missing or are only inconsistently found. The German, French and Italian speakers all manage the low, front vowel /æ/ although it is missing from these languages (no /ɛ /:/æ/ distinction). The French speaker devoices the final sound in 'things', but she pronounces it in 'please', and she uses the English approximant [ɹ] even though the French stereotypically pronounce /r/ as a uvular trill, [ʀ] or uvular fricative [ʁ]. We find variation not only in the various groups of speakers but also in the speech of individual speakers.

It is not surprising that individual non-native speakers vary in the degree to which they conform to stereotypes, i.e. in the strength of their accents. But since accents vary and a wide range of differences with respect to English all fall under the category

of ‘foreign accent’, we need a measure that takes many differences into account in assessing the strength of the foreign accent. We claim that the Levenshtein distance, introduced below, is appropriate in this respect because it yields a numerical measure representing the pronunciation difference per word, which may then be averaged over multiple words to obtain an aggregate measure of pronunciation difference.

2.2 The speakers

The speech accent archive contains transcribed speech samples (according to the International Phonetic Alphabet) from a large set of speakers. For all speakers it also contains their native language (people who are balanced bilinguals are excluded), other languages spoken, place of birth, age, gender, age of English onset (defined as the first exposure to sustained English language input), cumulative residence length in an English-speaking country, and learning style (i.e. naturalistic or academic).

In 2010, we extracted all available 989 transcribed samples from the Speech Accent Archive including speaker information. As there were only three speakers who were younger than 18, we excluded these from the dataset. Of all 986 adult speakers, 180 were native English and 115 of these native English speakers were born in the United States. The average age of all 986 speakers was 33.2 (SD: 13.1). There were more male than female speakers (555: 56% versus 431: 44%). The average age of English onset of the 806 non-native English speakers was 12.3 (SD: 7.4), while the mean residence length in an English-speaking country of these speakers was 7.7 years (SD: 11.7). A minority of the non-native English speakers (11.7%) learned English in a naturalistic (as opposed to an academic) setting.

We are aware that reading a paragraph of text may not be the best method to tap into pronunciation ability, as differences in reading ability may also affect the foreign accent (Piske, Mackay and Flege, 2001). However, the advantage of this approach is that a set of comparable text is obtained for every speaker, enabling a straightforward comparison.

3. Method

We wish to propose a technique to determine the degree of foreignness (i.e. foreign accent ratings) for any number of speech samples. As these ratings cannot easily be obtained behaviorally (by asking for several native speaker judgments per sample), we propose to use an automatic method to calculate them. Naturally, we need to validate the measure using native-speaker judgments, and this is the purpose of the present study. Assuming that our automated measurements are shown to be valid, we will then be in a position to use the automated measurements directly on larger sets of speech samples.

3.1. Automatically calculating foreignness ratings

The Levenshtein distance algorithm is able to calculate pronunciation distances between two transcribed strings by calculating the number of substitutions, insertions and deletions to transform one string of phonetic transcription symbols into the other (Levenshtein, 1965). For example, the Levenshtein distance between two accented pronunciations of the word Wednesday, [wɛ nzdeɪ] and [wɛ nəsde] is 3 as can be seen in the alignment below:

w	ɛ	n		z	d	e	ɪ
w	ɛ	n	ə	s	d	e	
			1	1			1

We note here that the calculation of Levenshtein distance automatically provides an alignment such as the one above in which corresponding segments are identified (Kruskal, 1999). Indeed the algorithm and its variants are often used primarily in order to identify corresponding elements. We further note that the Levenshtein distance is restricted to measuring differences in sequences of phonetic (or phonological segments). Suprasegmental information, including intonation, duration and tempo are not taken into account at all. So the Levenshtein distance is positioned to measure accent differences that are expressed segmentally, but not those that are reflected only suprasegmentally.

The Levenshtein distance has been successfully used for comparing pronunciations in dialectology (Kessler, 1995; Nerbonne et al., 1997; Nerbonne and Heeringa, 2010). Unfortunately, the standard Levenshtein distance algorithm is quite crude and only distinguishes same from different (i.e. substituting for a completely different sound, such as [u] for [i] is not distinguished from substituting for a more similar sound, such as [u] for [o]). To make the pronunciation comparison procedure more linguistically sensible, Wieling et al. (2009) proposed a method to incorporate (automatically obtained) sensitive sound segment distances in the Levenshtein distance algorithm and showed this approach improved the alignment quality significantly. The technique relies on the information-theoretic concept of pointwise mutual information (PMI)

and assigns smaller segment distances to segment pairs that align together frequently.[1] In a subsequent study, Wieling et al. (2012) showed the underlying sound (vowel) distances were linguistically sensible as they corresponded well to acoustic vowel distances, with correlations ranging from $r = 0.63$ to $r = 0.76$ for several datasets. Applying this method to our example alignment earlier yields the following associated costs:

w	ɛ	n			z	d	e	ɪ
w	ɛ	n	ə		s	d	e	
			0.031		0.020			0.030

In order to apply the PMI technique effectively, it is best that each segment occurs frequently (i.e. by including many words and speakers). This means that it is advantageous to reduce the number of different segments, which we do by ignoring diacritics, i.e., effectively treating [t], [t^{h}], [t̃], [t^{j}], etc. as the same segment. Naturally, this sacrifices some sensitivity in the measure, but without it, the frequencies of correspondences in alignment are too low to reliably obtain sensible segment distances. We obtain pronunciation distances per word using this (linguistically sensible) adaptation of the Levenshtein algorithm. As longer words are likely to vary more than shorter words, we divide the pronunciation distances by the alignment length. Pronunciation distances between two speakers can then simply be obtained by calculating the word pronunciation distances for all words and averaging these. Note that we tokenized the pronunciations of the Speech Accent Archive into separate words in order to support word-by-word comparison. We note that this procedure

[1] Other segment distances might be used, but as Laver (1994) notes, phonetics has not succeeded in providing general methods for measuring segment differences, except in the case of vowels.

respects the sandhi effects in pronunciation since we kept each word transcription exactly as it appeared, including whatever sandhi effects might be present. The tokenization procedure merely separated the transcription string into separate units corresponding to the words (as in the examples in Section 2.1).

To determine the foreignness rating of a speaker (with respect to U.S. English) we calculated the mean pronunciation distance between the transcribed speech sample of the foreign speaker and all 115 native U.S. English speakers in our dataset. Conceptually this can be interpreted as comparing the foreign pronunciation to the speech of the average U.S. English speaker.

3.2 Related work

McMahon et al. (2007) explicitly aim to measure the degree of accentedness in various forms of English world-wide, but they appeal to an algorithm that is not specified completely. They also criticize the work in the line of research presented here, but they get a number of crucial aspects wrong (Nerbonne, 2007). Nerbonne and Heeringa (2010) review a good deal of literature on the use of pronunciation distance measures focusing on measuring the similarity of pronunciation in the various dialects of a language, reporting on applications in more than a dozen languages, and noting that Gooskens and Heeringa (2004) show that Levenshtein distances correlate well ($r \approx 0.7$) with speakers' judgments of the dialect differences among Norwegian dialects.

Pairwise alignment methods, such as the Levenshtein algorithm, also enable computationally efficient well-performing multiple sequence alignment procedures (Prokić, Wieling and Nerbonne, 2009; List, 2012), which are important for historical

linguistics. Given the intimate relation between distance and alignment noted above, we interpret these results to indicate further that the Levenshtein distance is assaying pronunciation distance validly.

Wieling, Margaretha and Nerbonne (2012) worked with data sets from six different languages using the alignments provided by the Levenshtein algorithm to induce a measure of phonetic similarity in segments which they demonstrated correlated strongly with distances in formant space ($0.61 < r < 0.76$). Given the intimate relation between distance and alignment noted above, we interpret this result to indicate that the Levenshtein distance is assaying pronunciation distance validly.

Wieling, Prokić and Nerbonne (2009) introduced a validation of the Levenshtein distance using alignments, rather than perceived distances. They evaluated the pairwise alignments of Bulgarian dialect pronunciations, showing that their PMI-based Levenshtein method results in 97.5% accuracy when measured at the level of corresponding segments.[2]

Several non-dialectological studies have also successfully relied on the Levenshtein distance to measure pronunciation differences. Kondrak and Dorr (2004) used the Levenshtein distance to measure the pronunciation similarity of the names of proposed new drugs to existing ones. The goal was to avoid proposing names that patients, but also health personnel might easily confuse. Sanders and Chin (2009) use a version of the Levenshtein distance to measure the atypicalness of the speech of the

[2] The PMI-based approach we use here is slightly different from the original method proposed by Wieling et al. (2009). Wieling (2012; Ch. 2) discusses the modification and shows that it is slightly better than the original approach in terms of alignment quality (i.e. the accuracy improves to 97.7% for the Bulgarian data).

bearers of cochlear implants. In a study with aims similar to the present one Gooskens, Beijering and Heeringa (2008) showed that a Levenshtein distance based on segment distances derived from canonical spectrograms and normalized for length correlated extremely highly with intelligibility (r = -0.86). So we have every reason to be optimistic in proposing that the Levenshtein distance would be suitable to measure the strength of foreign accents in English pronunciations.

We turn now to the validation of pronunciation distances using judgments of native-likeness.

3.3 Validating automatically obtained foreignness ratings

Although several studies have used Levenshtein to measure pronunciation differences (see Section 3.2), there has been no validation of the method used to measure the strength of foreign accents to date. We therefore aim to fill that gap in this paper, and compare the computed Levenshtein distances to human native-likeness ratings.

We developed an online questionnaire in which native U.S. English speakers participants were presented with a randomly ordered subset of 50 speech samples from the Speech Accent Archive. We did not include all speech samples, as our goal was to obtain multiple native-likeness judgments per sample (to increase the reliability). For each speech sample, participants had to indicate how native-like each speech sample was. This question was answered using a 7-point Likert scale (ranging from 1: very foreign sounding to 7: native American English speaker). Participants were not required to rate all samples, but could rate any number of samples.

Via e-mail and social media we asked colleagues and friends to forward the online questionnaire to people they knew to be native U.S. English speakers. In addition, the online questionnaire was linked to in a post on *Language Log* by Mark Liberman.[3] Especially the latter announcement led to an enormous amount of responses. As a consequence, we replaced the initial set of 50 speech samples five times with a new set to increase the number of speech samples for which we could obtain native-likeness ratings. As there was some overlap in the native U.S. English speech samples present in each set (used to anchor the ratings), the total number of unique samples was 286, of which 272 were samples of non-native (U.S. or otherwise) English speakers.

4. Results

4.1. Validating the Levenshtein distance as a foreign accent rating

A total of 1143 native U.S. English participants filled in the questionnaire (658 men: 57.6%, and 485 women: 42.4%). Participants were born all over the United States, with the exception of the state of Nevada. Most people came from California (150: 13.1%), New York (115: 10.1%), Massachusetts (68: 5.9%), Ohio (66: 5.8%), Illinois (65: 5.7%), Texas (55: 4.8%), and Pennsylvania (54: 4.7%). The average age of the participants was 36.7 years (SD: 13.9) and every participant rated on average 41 samples (SD: 14.0).

In order to assess the consistency of the judgments we calculated Cronbach's alpha (Cronbach, 1951). The internal consistency was good with Cronbach's alpha equal to 0.853.

[3] http://languagelog.ldc.upenn.edu, May 19, 2012. "Rating American English Accents"

To find out how well the Levenshtein distance matched with the native-likeness ratings, we calculated the Pearson correlation *r* between the averaged ratings and the Levenshtein distances. For the 286 samples we found a correlation of $r = -0.78$, $p < 0.0001$. When using the log-transformed Levenshtein distances, the correlation was even stronger: $r = -0.81$, $p < 0.0001$. The direction is negative as the participants indicated how *native-like* each sample was, while the Levenshtein distance indicates how foreign a sample is. Figure 1 shows the scatterplot (including the trend line) of native-likeness as a function of the logarithm of the Levenshtein distance. Given these high correlations we may safely assume the automatically obtained Levenshtein distances are a valid means to assess foreign accent ratings in pronunciation.

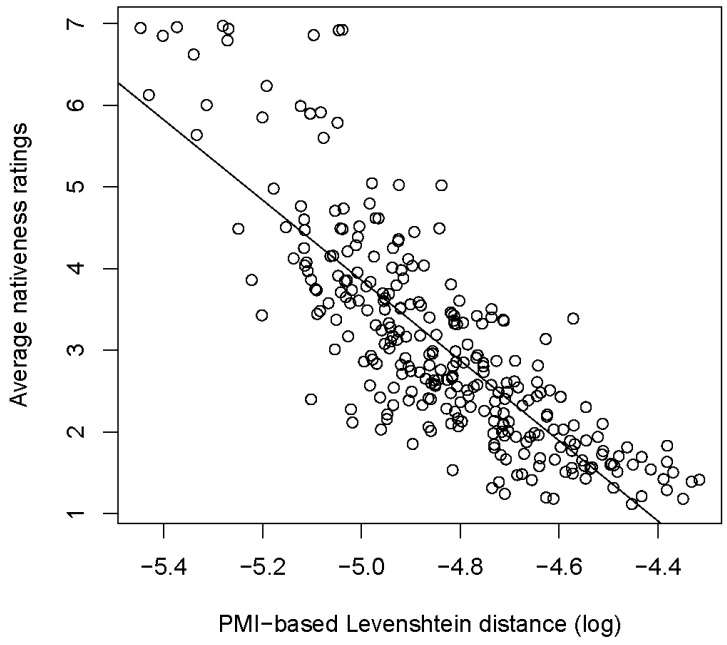

Figure 1. Logarithmically-corrected PMI-based Levenshtein distance as a predictor of mean native-likeness. See text for discussion.

Each point in Figure 1 pairs the Levenshtein measure of non-native-likeness with the mean judgment of the respondents to our questionnaire. Points far to the left represent very low Levenshtein distances which increase as one moves to the right on the *x*-axis. Vertically low points were judged to be very unlike native speech, and the similarity to native speech increases as one looks further up the *y*-axis. Examining the scatterplot more closely, we note that the cloud of points in the upper left of the graphs deviates from the trend line; for these points in the upper left, the Levenshtein distance tends to underestimate how native-like speech samples are when the differences to native pronunciation are judged to be small. As the number of native speakers in the dataset is much lower than the number of non-native speakers of English, the sound correspondences among native speakers will have a relatively low frequency and, consequently, a relatively high PMI segment distance, which may explain the higher distances. An alternative explanation might be that very natural suprasegmental qualities might "compensate" for segmental differences where these are small.

If the measure correlates with human judgments at the level of $r = 0.8$, then it accounts for a good deal, but not all of the variance in the comparison ($r^2 = 0.64$). There are two important candidates to explain the remaining 36%. The first is the suprasegmental information, which we systematically ignored (see above). The second is the transcription process. While the transcription quality of the Speech Accent Archive seems excellent, we do not know of measures of transcriber agreement (Weinberger and Kunath, 2011), and the fact remains that transcription is a very difficult and error-prone task.

5. Conclusions and discussion

We used a large set of transcribed data from non-native speakers of English who read the same paragraph aloud (Weinberger and Kunath 2011), and used the Levenshtein distance to measure how much the non-native speech differed from native American English speech. In particular we used a version of the Levenshtein distance which employs automatically induced segment distances, introduced by Wieling et al. (2009), and normalized for alignment length. We collected judgments of native-likeness from over 1,100 native speakers and showed that their mean judgments correlated strongly with the logarithmically corrected computational measure (r = -0.81). This shows that the Levenshtein measure may serve as a proxy for human judgments of non-native-likeness, allowing us to study this phenomenon in a replicable way without incurring the expense of human judgments.

One further task is clear, namely to investigate what sorts of factors predict non-native-likeness while taking into account a large group of non-native speakers with various language backgrounds. A second task would be to investigate refinements of the Levenshtein distance in order to develop a technique even better able to gauge pronunciation differences, perhaps focusing on ways to include both segmental and supra-segmental information or on ways of incorporating the fine-grained information present in the diacritics.

Acknowledgements

We are very grateful to Mark Liberman for his post on Language Log inviting native U.S. American English speakers to rate the speech samples for the native-likeness.

References

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3): 297-334.

Gooskens, C. and W. Heeringa (2004). Perceptive Evaluation of Levenshtein Dialect Distance Measurements using Norwegian Dialect Data. *Language Variation and Change*, 16(3): 189-207.

Gooskens, C., K. Beijering and W. Heeringa (2008). Phonetic and lexical predictors of intelligibility. *International Journal of Humanities and Arts Computing*, 2(1-2): 63-81.

Kessler, B. (1995). Computational dialectology in Irish Gaelic. In: *Proc. 7th Seventh European ACL*, Dublin, 60-66.

Kondrak, G. and B.J. Dorr (2004). Identification of Confusable Drug Names: A New Approach and Evaluation Methodology, *Proc. of COLING*, Geneva. 952-958.

Kruskal, J. [1983] (1999). An Overview of Sequence Comparison. In D. Sankoff and J. Kruskal (Eds.) *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*, 1–44. Reprinted, with a foreword by J. Nerbonne, Stanford, CA: CSLI Publications.

Laver, J., (1994). *Principles of Phonetics*. Cambridge University Press, Cambridge

Levenshtein, V. (1965). Binary codes capable of correcting deletions, insertions and reversals. *Doklady Akademii Nauk SSSR*, 163:845–848. In Russian.

List, J. M. (2012). Multiple Sequence Alignment in Historical Linguistics. A Sound Class Based Approach. *Proc. ConSOLE XIX*, 241–260. http://media.leidenuniv.nl/legacy/console19-proceedings-list.pdf (8 Apr. 2013, date last accessed).

McMahon, A., P. Heggarty, R.McMahon, and W.Maguire (2007) The sound patterns of English: Representing phonetic similarity. *English Language and Linguistics*, 11(1): 113-142. DOI: 10.1017/S1360674306002139.

Nerbonne, J. (2007). Review of A. McMahon & R. McMahon *Language Classification by the Numbers.* Oxford: OUP. 2005. *Linguistic Typology*, 11: 425-436

Nerbonne, J. and W. Heeringa (1997). Measuring Dialect Distance Phonetically. In: J. Coleman (Ed.) *Workshop on Computational Phonology*. (SIGPhon) ACL: Madrid, 11–18.

Nerbonne, J. and W. Heeringa (2010). Measuring Dialect Differences. In: J. E. Schmidt and P. Auer (Eds.) *Language and Space: Theories and Methods* in series *Handbooks of Linguistics and Communication Science*. Berlin: Mouton De Gruyter, Chap. 31, 550-567.

Piske, T., I.R.A. MacKay, and J. E. Flege. (2001). "Factors affecting degree of foreign accent in an L2: A review." *Journal of Phonetics* 29(2): 191-215.

Prokić, J., M. Wieling and J. Nerbonne (2009). Multiple Sequence Alignments in Linguistics. In L. Borin & P. Lendvai (Eds.) *Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities and Education* (LaTeCH - SHELT&R 2009) Workshop at the 12th EACL. Athens. 18-25.

Sanders, N. C. and Chin, S. B. (2009). Phonological distance measures. *Journal of Quantitative Linguistics*, 43: 96–114

Weinberger, S. H. and S. A. Kunath (2011). The Speech Accent Archive: towards a typology of English accents. In: J. Newman, R.H. Baayen and S. Rice (Eds.) *Corpus-based Studies in Language Use, Language Learning, and Language Documentation*. Rodopi: Amsterdam/New York. 265-281. (Series: *Language and Computers*, 73).

Martijn Wieling (2012). *A Quantitative Approach to Social and Geographical Dialect Variation*. PhD dissertation, University of Groningen.

Wieling, M., Margaretha, E., and J. Nerbonne (2012). Inducing a measure of phonetic similarity from dialect variation. *Journal of Phonetics*, 40(2): 307–314.

Wieling, M., Prokić, J., and J. Nerbonne (2009). Evaluating the pairwise alignment of pronunciations. In: L. Borin & P. Lendvai (Eds.) *Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities and Education* (LaTeCH - SHELT&R 2009) Workshop at the 12th EACL. Athens. 26–34.