Computational Linguistics in the Netherlands Journal 1 (2011) 109-118 Submitted 7/2011; Published 12/2011

Inducing phonetic distances from dialect variation

Martijn Wieling M.B.WIELING@RUG.NL
Eliza Margaretha ELIZA.MARGARETHA@GMAIL.COM
John Nerbonne J.NERBONNE@RUG.NL
Department of Humanities Computing
University of Groningen

Abstract

In this study we attempt to derive phonetic distances from alternative dialectal pronunciations used in different geographical varieties. We use two dialect atlases each containing the phonetic transcriptions of the same set of words at hundreds of sites. We collect the sound correspondences through alignment with the Levenshtein distance algorithm, and then apply an information-theoretic measure, pointwise mutual information, assigning smaller segment distances to segments which frequently correspond. We iterate alignment and information-theoretic distance assignment until both stabilize and we evaluate the quality of the phonetic distances obtained by comparing them to acoustic vowel distances. For both Dutch and German, we find strong correlations between the induced phonetic distances and the acoustic distances, illustrating the usefulness of the method in deriving valid phonetic distances from dialectal pronunciations.

1. Introduction

In this study we attempt to automatically derive phonetic segment distances on the basis of how frequently the segments correspond in different (dialectal) pronunciations of the same words. We evaluate the success of the attempt by comparing vowel distances we derive to independent acoustic characterizations.

There are several perspectives which motivate this work. First, we have conducted a large number of studies using the Levenshtein distance (Levenshtein 1965) to assay pronunciation differences among dialects (Nerbonne and Heeringa 2009). The Levenshtein distance is implemented using an algorithm which sums individual sound segment distances to determine the distance between two pronunciations. For that reason we have experimented with a large number of segment distance measures, but none have been shown to improve (much) on the very simple, binary measure which distinguishes only identical and non-identical segments (Heeringa 2004, pp. 27-120, 186). Recently, however, Wieling et al. (2009) and Wieling and Nerbonne (2011) found that using automatically derived segment distances (using the procedure explained in Section 3.1) improved alignment quality considerably.[1] The purpose of this study is to show that these segment distances are linguistically sensible.

Second, the improved alignments, which make use of the induced segment distances are in turn useful in (automatically) identifying the sound correspondences which historical linguistics relies on (Prokić 2010, Ch. 6). Indeed, historical examination normally relies

1. For completeness, we have included previously published results illustrating the improved performance of this algorithm with respect to alignment quality in Section 4.1.

on detecting regular sound correspondences. These need not be similar sounds, naturally, but the procedure we describe below generalizes to cases in which correspondences are less phonetically similar.

Third, as Laver (1994, p. 391) notes, there is no widely accepted procedure for determining phonetic similarity, nor even explicit standards: "Issues of phonetic similarity, though underlying many of the key concepts in phonetics, are hence often left tacit." We wish to add a means of using distributions of variation in pronunciation to other techniques for detecting and determining similarity.

1.1 Related work

Kernighan et al. (1990) examined the problem of suggesting intended spellings once a misspelling has been detected. They derive the posterior chances of alternative candidates that differed in just one insertion, deletion, substitution or transposition by initially assuming uniform chances for each operation and then updating estimates empirically based on words that were spelled incorrectly. We share with their work the attempt to derive weights for edit operations empirically. They evaluated their work by noting how often they could select the intended spelling from a list of alternatives (provided by UNIX *spell*). We shall embed the weights we derive in a version of the Levenshtein algorithm and we evaluate the results by examining the improvement in alignment quality and also by checking the correlation of the weights we find with acoustically determined distances.

Wieling et al. (2007a) used a Pair Hidden Markov Model to align dialect pronunciations and they evaluated their work by checking the correlation of the emission probabilities for pairs of sounds with acoustically determined weights. Wieling et al. (2009) also evaluated this technique with respect to alignment quality. While the alignment performance was similar to the approach illustrated in the present paper, the Pair Hidden Markov Model was computationally much more expensive and less transparent in its errors (Wieling et al. 2007a).

2. Material

2.1 Dialect pronunciations

In this study we derive phonetic distances for two data sets, a Dutch and a German dialect data set. The Dutch dialect data set contains phonetic transcriptions of 562 words in 613 locations in the Netherlands and Flanders. Wieling et al. (2007b) selected the words from the Goeman-Taeldeman-Van-Reenen-Project (GTRP; Goeman and Taeldeman, 1996) specifically for an analysis of pronunciation variation in the Netherlands and Flanders. The German data set contains phonetic transcriptions of 201 words in 186 locations collected from the *Phonetischer Atlas der Bundesrepublik Deutschland* (Göschel 1992) and was analyzed and discussed in detail by Nerbonne and Siedle (2005).

2.2 Acoustic vowel measurements

For Dutch, we used vowel frequency (Hertz) measurements of 50 male (Pols et al. 1973) and 25 female (van Nierop et al. 1973) speakers. In line with Wieling et al. (2007a), we only included vowels which are pronounced as monophthongs in standard Dutch, yielding

measurements for nine vowels: /i, ɪ, y, ʏ, ɛ, a, ɑ, ɔ, u/. For German, we used vowel frequency measurements of 69 male and 58 female speakers (Sendlmeier and Seebode 2006) for fourteen vowels /i, ɪ, y, ʏ, e, ɛ, a, o, ɔ, u, ʊ, ʌ, ɘ, ə/. For both languages, we averaged the mean frequencies of men and women in order to obtain a single set of frequencies.

3. Methods

In a nutshell, the procedure we describe first aligns different dialect pronunciations, using a binary, same-different measure of segment difference. We keep track of how often each sound correspondence occurs, applying an information-theoretic measure of association strength, which in turn is used to provide a new estimation of the segment distance. We then re-align the dialect pronunciations, this time using the newly acquired segment distances. The process is repeated until the segment distances (and alignments) stabilize. In the following section, we describe these steps in more detail.

3.1 Obtaining sound distances based on dialect pronunciations

We automatically determine the sound segment distances on the basis of their co-occurrence in different dialectal pronunciations of the same word. To identify co-occurring sounds we generate alignments based on the Levenshtein distance (Levenshtein 1965) which minimizes the number of insertions, deletions and substitutions to transform one string into the other.

For example, the Levenshtein distance between two Dutch variants of the word 'to bind', [bɪndən] and [bɛində], is 3:

bɪndən	insert ɛ	1
bɛɪndən	subst. i/ɪ	1
bɛindən	delete n	1
bɛində		
		3

The corresponding alignment is:

b		ɪ	n	d	ə	n
b	ɛ	i	n	d	ə	
	1	1				1

The regular Levenshtein distance does not distinguish vowels and consonants and therefore may align a vowel with a consonant. To enforce linguistically sensible alignments we added a syllabicity constraint such that vowels are not aligned with consonants. This is the only information about phonetic content made available to the (basic) system.

Note that each point in the alignment in which non-identical sounds are aligned is assigned a cost of 1. More sophisticated versions of the Levenshtein distance can make use of more discriminating costs. In fact, one can define a table of segment distances and use these in the algorithm.

It is, however, difficult to obtain (complete) segment distance tables to use in conjunction with the Levenshtein algorithm, in particular if one's goal is to characterize (nearly) all the distinctions made in dialect atlases. For example, *The Linguistic Atlas of the Middle*

and South Atlantic States (Kretzschmar 1994) distinguishes over 1100 different vowels (combinations of base segments with one or more diacritics) and nearly 1700 different segments in total. Nevertheless, Heeringa (2004) experimented with three different segment distance tables, two feature-based tables — one based on Chomsky and Halle's *Sound Pattern of English* (Chomsky and Halle 1968), the other on Almeida and Braun's system designed to assess transcription accuracy (Almeida and Braun 1986) — as well as a system derived from curve distance in canonical spectrograms (Heeringa 2004, Ch. 4). The final results could not be shown to be superior to the binary system of differences, however, at least not when validated in the aggregate as correlates of dialect speakers' judgments of how "different" other varieties sound (Heeringa 2004, p. 186). Our inductive procedure seeks to bypass the need for an expert's specification of a segment distance table.

Since we are looking at dialectal pronunciations which are reasonably similar to each other, it is conceivable that similar sounds like [i] and [y] will co-occur more frequently than more distant sounds such as [a] and [i].

Pointwise mutual information (PMI; Church and Hanks, 1990) was used by Wieling et al. (2009) to determine the distance between every pair of sounds on the basis of their relative frequency of co-occurrence. Wieling et al. (2009) found that using the Levenshtein distance with PMI-based sound distances resulted in improved alignments of Bulgarian dialectal pronunciations compared to using the Levenshtein algorithm with a syllabicity constraint (which does not distinguish varying levels of sound similarity).

The PMI approach consists of obtaining initial string alignments for a corpus of dialectal material by using the Levenshtein algorithm with syllabicity constraint. After the initial run, the substitution cost of every sound segment pair is calculated according to the PMI procedure assessing the statistical dependence between the two sounds:

$$\mathrm{PMI}(x,y) = \log_2\left(\frac{p(x,y)}{p(x)\,p(y)}\right)$$

Where:

- $p(x,y)$ is estimated by calculating the number of times sound segments x and y occur at the same position in two aligned pronunciations X and Y, divided by the total number of aligned segments (i.e. the relative occurrence of the aligned sound segments x and y in the whole data set).
- $p(x)$ and $p(y)$ are estimated as the number of times sound segment x (or y) occurs, divided by the total number of segment occurrences (i.e. the relative occurrence of sound segments x or y in the whole data set). Dividing by this term normalizes the correspondence frequency with respect to the frequency expected if x and y are statistically independent.

Positive PMI values indicate that sounds tend to co-occur in correspondences (the greater the PMI value, the more two sounds tend to co-occur), while negative PMI values indicate that sounds do not tend to co-occur in correspondences. Sound distances (i.e. sound segment substitution costs) are generated by subtracting the PMI value from 0 and adding the maximum PMI value (to ensure that the minimum distance is 0).

Following Wieling and Nerbonne (2011), we ignore pairs of identical sounds, as this modification improved the quality of the Bulgarian pronunciation alignments with respect to

the original approach of Wieling et al. (2009). From the perspective of string transformation, there can be no cost associated with retaining a segment, and from the perspective of alignment, no cost accrues to aligning identical sounds.

After the new sound segment substitution costs have been calculated for the first time, the pronunciations are aligned anew based on the adapted sound distances. This process is repeated until the pronunciation alignments and sound distances remain constant. How well these final sound distances correspond with acoustic sound distances is discussed in Section 4.

3.2 Calculating acoustic distances

To obtain the acoustic distances between vowels, we calculate the Euclidean distance of the formant frequencies (in Bark). As our perception of frequency is non-linear, calculating the Euclidean distance on the basis of Hertz values would not weigh the first formant enough. We therefore convert the Hertz frequencies to Bark scale (Traunmüller 1990) in better keeping with human perception.

4. Results

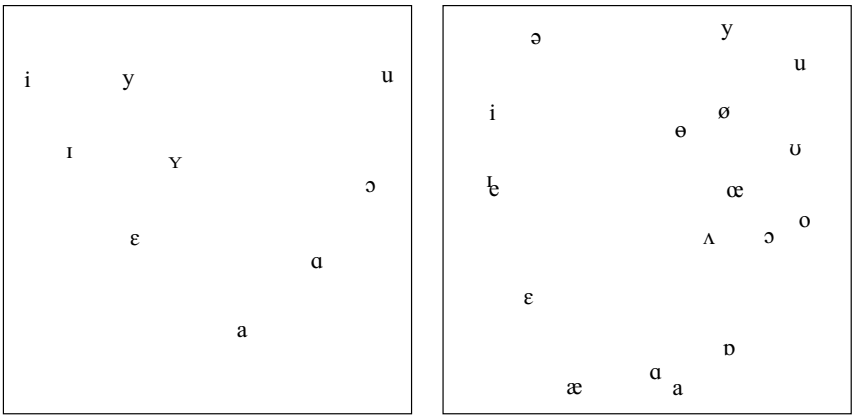
The Dutch dialect pronunciation data set contains 26 different vowels, some of which occur relatively infrequently. To obtain a reliable set of vowel distances, we excluded all vowels (8) having a frequency lower than one percent of the maximum vowel frequency. The final vowel set consisted of 18 vowels: /a,ɑ,ɒ,ʌ,æ,e,ɛ,i,ɪ,y,o,ɔ,u,ʊ,ɵ,œ,ø,ə/.

The German dialect pronunciation data set contains 28 vowels, of which 7 were excluded as they had a frequency lower than one percent of the maximum vowel frequency. The final German vowel set consisted of 21 vowels: /a,ɑ,ɒ,ʌ,ɐ,æ,e,ɛ,i,ɪ,y,ʏ,o,ɔ,u,ʊ,ɯ,ɵ,œ,ø,ə/.

Given a matrix of vowel distances, we can use multidimensional scaling (MDS; Togerson, 1952) to place each vowel at the optimal position relative to all other vowels in a two-dimensional plane. Figure 1(a) shows the relative positions of the Dutch vowels on the basis of their acoustic distances (the complete variance is visualized in the two dimensions), while Figure 1(b) shows the relative positions of the Dutch vowels based on their PMI-based distances (76% of the variance is visualized in two dimensions). Similarly, Figure 2(a) shows the relative positions of the German vowels based on their acoustic distances (the complete variance is visualized), while Figure 2(b) shows the positioning of the German vowels on the basis of their PMI-based distances (70% of the variance is visualized). Note that the number of vowels for which acoustic measurements were available was lower than the number of vowels distinguished in the transcribed dialect pronunciations.

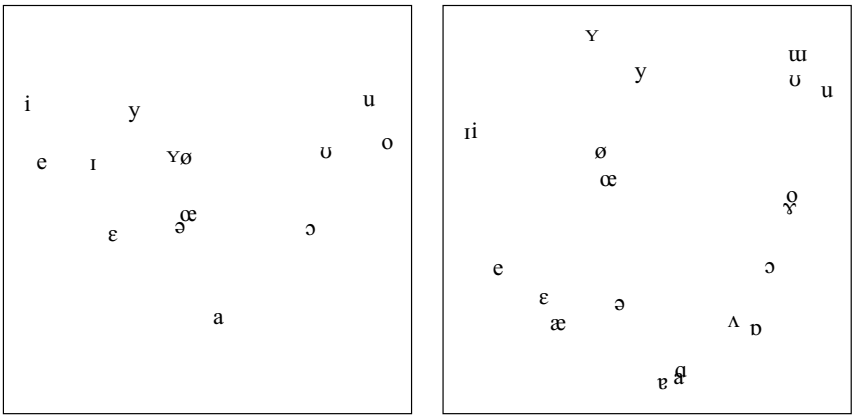
It is clear that the visualizations on the basis of the acoustic distances resemble the IPA vowel chart (shown in Figure 3) quite nicely. The visualizations on the basis of the PMI distances are less striking. We certainly can identify many resemblances with the IPA vowel chart when examining the PMI-based graphs more closely, however. The positions of [i], [u], [a] and similar sounds are quite acceptable, considering the distances are based *only* on how frequently the sounds align in dialect data.

In the Dutch PMI-based visualization, however, the position of the [ə] (schwa) deviates significantly from the position on the basis of the acoustic distances. Investigating the alignments revealed that the schwa was frequently deleted (i.e. aligned against a gap) and



(a) Acoustic distance visualization (b) PMI distance visualization

Figure 1: Relative positions of Dutch vowels based on their acoustic (a) and PMI distances (b). The visualization in (a) captures 100% of the variation in the original distances, while the visualization in (b) captures 76% of the variation in the original distances.



(a) Acoustic distance visualization (b) PMI distance visualization

Figure 2: Relative positions of German vowels based on their acoustic (a) and PMI distances (b). The visualization in (a) captures 100% of the variation in the original distances, while the visualization in (b) captures 70% of the variation in the original distances.

this resulted in relatively high distances between the schwa and the other vowels (which were deleted less frequently) compared to the other distances. Excluding the schwa increased the ability to visualize the relations between the vowels adequately in two dimensions: the explained variance increased from 76% to 85%. While the schwa was positioned better in Figure 2(b), we found that the schwa was the most frequently deleted sound in the German data set. Consequently, excluding the schwa from the visualization increased the explained

variance of the two dimensional visualization from 70% to 82%. A second striking deviation is the position of the [y] (and [ʏ] for the German data set), for which we have no immediate explanation.

Besides looking at the similarities between the multidimensional scaling results, we can also measure how well the PMI distances correspond with the acoustic distances for sounds present in both sets. For the Dutch data, the correlation between the acoustic and PMI distances was $r = 0.657$ ($p < 0.001$). Note that the deviating position of the schwa did not have an effect on this correlation, as there were no acoustic measurements for this sound (see Figure 1(a)). For the German data, the correlation was $r = 0.630$ ($p < 0.001$). However, when the schwa was excluded, the correlation increased to $r = 0.785$ ($p < 0.001$).[2]

4.1 A note on alignment quality

Wieling et al. (2009) evaluated the initial PMI-based Levenshtein algorithm with respect to several other algorithms using a Bulgarian dialect data set. Wieling and Nerbonne (2011) reported that when ignoring pairs of identical sounds in the original PMI procedure the alignments improved, but they did not report the exact improvement.

To illustrate the performance of the PMI-based Levenshtein algorithm at the alignment level, Table 1 shows the number of misaligned segments and non-identical alignments with respect to the (manually corrected) gold standard alignments of the Bulgarian dialect data set. A detailed description about this data, the creation of the gold standard alignments as well as the procedure to measure the number of segment errors is given by Wieling et al. (2009).

The regular Levenshtein algorithm employs a binary same-different measure and does not align vowels with consonants. The initial PMI-based Levenshtein algorithm included pairs of identical sounds in the counts necessary for the PMI calculation (Wieling et al. 2009), while the improved PMI-based Levenshtein algorithm employed in the current study ignored these (Wieling and Nerbonne 2011). Table 1 clearly illustrates that the PMI-based Levenshtein algorithms significantly outperformed the regular Levenshtein algorithm, and the improved PMI-based Levenshtein algorithm slightly but significantly outperformed the initial PMI-based Levenshtein algorithm.

Unfortunately we did not have gold standard alignments for either the Dutch or the German dialect data set, but we have no reason to believe that results based on these data sets would show a different pattern. We therefore conclude that the PMI-based Levenshtein algorithm as outlined and evaluated here is highly suitable to obtain good alignments with a strong linguistic basis.

5. Discussion and conclusion

Based on the results discussed in the previous section, we conclude that we are able to characterize the phonetic distance between segments to a surprising extent on the basis of the distribution of the segment's pronunciation variants among closely related varieties. Since we tested this conclusion based on an acoustic measure for those segments where a

2. We assessed the significance of the correlation coefficients by using the Mantel test (Mantel 1967), as our sound distances are not completely independent.

Algorithm	Segment errors	Alignment errors (%)
Regular Levenshtein	490,703	191,674 (5.52%)
PMI-based Levenshtein (initial)	399,216	156,440 (4.50%)
PMI-based Levenshtein (improved)	**387,488**	**152,808 (4.40%)**

Table 1: Comparison to gold standard alignments. All differences are significant ($p < 0.01$).

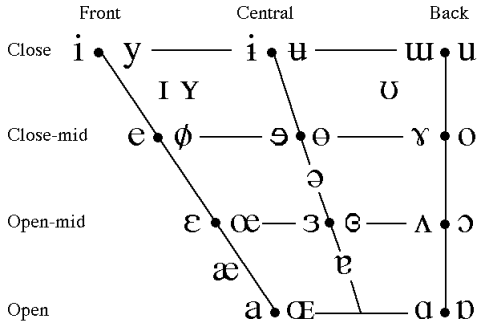


Figure 3: Vowel chart of the International Phonetic Alphabet

measure is well established, we may conjecture that the segment distances also correlate well in those cases for which we still lack appropriate validating material.

The level of correlation was similar in the two independent dialect data sets, an encouraging indication that the relation between functioning as an alternative pronunciation and being similar in pronunciation is neither accidental nor trivial. However, as German and Dutch are similar languages, it would be useful to investigate dialects from more distantly related languages.

The opportunity to exploit phonetic segment distances in string alignment and string distance algorithms will allow us to assess word (string) distances more accurately and to improve pronunciation alignments. This is valuable in dialectometry and also in historical linguistics where the determination of regular sound correspondences is important.

Since we evaluated the quality of the automatically obtained segment distances with respect to acoustic vowel distances, it might seem that we could just have used these instead in our alignment procedure. There are two problems with this approach. First, acoustic sound distances are only available for vowels as it is currently unclear how to obtain these for consonants. Second, acoustic vowel measurements might not always be readily available for every language. As our method automatically generates (acoustically sensible) sound distances, our method does not have this restriction.

An intriguing aspect of this work is that distributions (of variant pronunciations) contain enough information to gauge content (i.e. phonetic similarity) to some extent. The only phonetic content made available to the algorithm was the distinction between vowels and consonants, and yet the algorithm could assign a phonetic distance to all pairs of vowel segments in a way that correlates strongly with acoustic similarity.

Acknowledgments

We thank Peter Kleiweg for implementing the PMI procedure in the L04 package which was used to generate the vowel distance visualizations. Mark Liberman discussed the ideas in this paper with us generously. We also thank two anonymous reviewers whose comments helped to improve this paper.

References

Almeida, Almerindo and Angelika Braun (1986), 'Richtig' und 'Falsch' in phonetischer Transkription: Vorschläge zum Vergleich von Transkriptionen mit Beispielen aus deutschen Dialekten, *Zeitschrift für Dialektologie und Linguistik* **LIII** (2), pp. 158–172.

Chomsky, Noam A. and Morris Halle (1968), *The Sound Pattern of English*, Harper and Row, New York.

Church, Kenneth W. and Patrick Hanks (1990), Word association norms, mutual information, and lexicography, *Computational Linguistics* **16** (1), pp. 22–29.

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

Göschel, Joachim (1992), Das Forschungsinstitut für Deutsche Sprache "Deutscher Sprachatlas". Wissenschaftlicher Bericht, Das Forschungsinstitut für Deutsche Sprache, Marburg.

Heeringa, Wilbert (2004), *Measuring Dialect Pronunciation Differences using Levenshtein Distance*, PhD thesis, Rijksuniversiteit Groningen.

Kernighan, Mark, Kenneth Church, and William Gale (1990), A spelling correction program based on a noisy channel model, *Proceedings of the 13th conference on Computational linguistics*, Vol. 2, Association for Computational Linguistics, pp. 205–210.

Kretzschmar, William A., editor (1994), *Handbook of the Linguistic Atlas of the Middle and South Atlantic States*, The University of Chicago Press, Chicago.

Laver, John (1994), *Principles of Phonetics*, Cambridge Univeristy Press, Cambridge.

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

Mantel, N. (1967), The detection of disease clustering and a generalized regression approach, *Cancer Research* **27**, pp. 209–220.

Nerbonne, John and Christine Siedle (2005), Dialektklassifikation auf der Grundlage aggregierter Ausspracheunterschiede, *Zeitschrift für Dialektologie und Linguistik* **72**, pp. 129–147.

Nerbonne, John and Wilbert Heeringa (2009), Measuring dialect differences, *in* Schmidt, Jürgen Erich and Peter Auer, editors, *Theories and Methods*, Language and Space, Mouton De Gruyter, Berlin, pp. 550–567.

Pols, Louis, H. Tromp, and R. Plomp (1973), Frequency analysis of dutch vowels from 50 male speakers, *The Journal of the Acoustical Society of America* **43**, pp. 1093–1101.

Prokić, Jelena (2010), *Families and Resemblances*, PhD thesis, Rijksuniversiteit Groningen.

Sendlmeier, Walter and Julia Seebode (2006), Formantkarten des deutschen Vokalsystems. TU Berlin, `http://www.kgw.tu-berlin.de/forschung/Formantkarten` (accessed: November 1, 2010).

Togerson, W. (1952), Multidimensional scaling. I. Theory and method, *Psychometrika* **17**, pp. 401–419.

Traunmüller, H. (1990), Analytical expressions for the tonotopic sensory scale, *The Journal of the Acoustical Society of America* **88**, pp. 97–100.

van Nierop, D., L. Pols, and R. Plomp (1973), Frequency analysis of Dutch vowels from 25 female speakers, *Acoustica* **29**, pp. 110–118.

Wieling, Martijn and John Nerbonne (2011), Measuring linguistic variation commensurably, *Dialectologia* **Special Issue II: Production, Perception and Attitude**, pp. 141–162.

Wieling, Martijn, Jelena Prokić, and John Nerbonne (2009), Evaluating the pairwise alignment of pronunciations, *in* Borin, Lars and Piroska Lendvai, editors, *Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education*, pp. 26–34.

Wieling, Martijn, Therese Leinonen, and John Nerbonne (2007a), Inducing sound segment differences using Pair Hidden Markov Models, *in* Nerbonne, John, Mark Ellison, and Greg Kondrak, editors, *Computing and Historical Phonology: 9th ACL Special Interest Group for Morphology and Phonology*, pp. 48–56.

Wieling, Martijn, Wilbert Heeringa, and John Nerbonne (2007b), An aggregate analysis of pronunciation in the Goeman-Taeldeman-Van Reenen-Project data, *Taal en Tongval* **59**, pp. 84–116.