Measuring Dialect Pronunciation Differences using Levenshtein Distance

Heeringa, Wilbert Jan

Chapter 10

Conclusions and future prospects

10.1 Conclusions

The goal of this thesis is to explore whether Levenshtein distance could be a useful tool for measuring dialect word pronunciation distances, and thus for measuring dialect distances. Since we want to be able to compare the Levenshtein distance with the corpus frequency method and the frequency per word method, the two latter methods were also involved in the research. In Section 2.4 we concluded that the frequency per word method is methodologically better than the corpus frequency method, and the Levenshtein distance is methodologically better than the frequency per word method.

When attempting to quantify distances in pronunciation between dialects, we need to determine the relations between different speech segments. For this purpose we investigated discrete representations of segments (Chapter 3) and acoustic representations of segments (Chapter 4). The *phone* representation is the least discriminating discrete representation. Two segments are equal or unequal. Using feature representations gradual segment distances can be obtained. We examined the feature systems of Hoppenbrouwers & Hoppenbrouwers (H & H), Vieregge & Cucchiarini (V & C) and Almeida & Braun (A & B). When correlating segment distances obtained on the basis of these systems we found that the systems of V & C and A & B appeared to be most similar, although the correlations between these two systems were not significantly stronger than comparable correlations between any other pair of systems. Even these systems are different, however, as indicated by the rather low, but significant correlations. The use of the different systems will yield different results. We also investigated different metrics for comparing feature histograms (used in frequency-based methods) and feature bundles (used in Levenshtein distance). The correlations between the Manhattan metric and the Euclidean metric were stronger than the corresponding correlations between any other pair of metrics. Partly they were also significantly stronger. This indicates that results obtained on the basis of

Manhattan distance will not strongly differ from results obtained on the basis of Euclidean distance. The Pearson correlation coefficient appeared to be rather different from the two other metrics.

The acoustic segment representations we examined were the Barkfilter representation, the cochleagram representation and the formant track representation. These representations are more perceptually oriented than the commonly used type of spectrogram which has a Hertz-scale. On the basis of the distances obtained by the different representations we applied multidimensional scaling and scaled the distances to two dimensions. For the vowels we obtained a vowel quadrilateral (Barkfilter and cochleagram) or vowel triangle (formant tracks) and for the consonants a distinction between the different manners of articulation. These were reasonable results, although they were based on only two speakers. We conclude that the use of acoustic representations is useful, but recommend future work to verify the conclusion on the basis of more speakers, and if necessary to refine the acoustic processing. When comparing the present results of the different representations, we found that the Barkfilter results and the cochleagram results correlate significantly more strongly than the other pairs of representations. The formant track results appeared to be more different, so the use of the formant track representation will yield significantly different results than when using the Barkfilter or cochleagram representation.

We compared the acoustic representations with the feature-based representations. For vowels we found that the Barkfilter distances and the cochleagram distances correlate strongest with the A & B distances, but correlations with other feature systems were not significantly weaker. The formant track distances correlate strongest with the V & C distances. However the correlations were mostly not significantly higher than comparable ones of other feature systems. For consonants the Barkfilter distances and the cochleagram distances correlate strongest with the V & C distances, but only significantly stronger than with the A & B distances. The formant track distances correlate strongest with the V & C distances (RND) or H & H distances (IPA). The correlation coefficients were only significantly higher than the comparable ones of A & B. The correlations between acoustic distances and feature-based distances were not extremely high, although they were mostly significant. Therefore, both types of segment representations were considered in validation work.

In Chapter 5 we described Levenshtein distance. The Levenshtein distance of two word pronunciations is equal to the set of operations with the least cost which changes the one pronunciation into the other. We used insertions, deletions and substitutions. Future work may be to add the swap operation and to find the correct weight for this operation. The distance between two dialects is equal to the average word distance. We apply the Levenshtein distance to transcriptions where operations are applied to the transcription segments, and to recordings where operations are applied to spectra or formant bundles.

Once the distances between dialects are obtained, the dialects can be classified. Cluster analysis and multidimensional scaling are explained in Chapter 6. We examined several cluster methods and found that *Unweighted Pair Group Method using Arithmetic averages* to be methodologically superior to the other methods. We examined three different multidimensional scaling algorithms and found *Kruskal's Non-metric Multidimensional Scaling* preferable, since the results of this procedure represent the original distances with the greatest fidelity.

In Chapter 7 we validated different versions of the corpus frequency method, the frequency per word method and the Levenshtein distance on the basis of 15 Norwegian varieties. We used recordings and transcriptions of the fable 'The North Wind and the Sun', in Norwegian: 'Nordavinden og sola' (NOS). The data was compiled by Jørn Almberg. In the text we found 58 different words. In advance consistency was checked for the word-based methods. We calculated Cronbach's α values and found that the 58 words were enough to obtain reliable results. For one particular Levenshtein variant we found that the use of only 25 words gave already an acceptable degree of consistency ($\alpha = 0.70$).

Subsequently, for both transcription-based methods and recording-based methods we compared the measurements with the results of a perception experiment in which dialect speakers themselves judge the distances between the varieties. Examining the transcription-based methods it appeared that results obtained by methods using phones and the logarithmic Levenshtein distances using acoustic representations correlate most strongly with the perceptual distances. At first glance, this may be a partly unexpected outcome, but the methods share the property that small segment distances are relatively heavily weighted, which is perhaps also the case in perception. Among the feature representations, the H & H system yields the best results. Among the acoustic representations we found the Barkfilter representation better than the other two representations, but only when using the linear Levenshtein distance. Furthermore, we found that the use of 4 length gradations is preferable to 2 length gradations in general.[1] The computations did not clarify whether two-segmental representations of diphthongs are better than one-segmental representations, or the other way round. When representing speech segments by features, Manhattan is mostly preferable when using the corpus frequency method, and Euclidean is the better candidate when applying word-based methods. Using the Euclidean metric larger differences are weighted relatively more heavily than smaller differences. When using the corpus frequency method, dialect distances are measured with the metrics. Using the frequency per word method and Levenshtein distance, respectively word distances and segment distances are calculated with these metrics. This indicates that on the highest level (comparison of dialects) differences should

[1]When using 4 length gradations extra-short, short, half-long and long are represented by multiplying segments in the transcription. When using 2 length gradations only extra-short and non-extra-short are represented by multiplying segments in the transcription.

be weighted equally, but on the lower levels (comparison of words or segments) larger differences should be weighted relatively more heavily than smaller ones. The best method is a variant of the transcription-based Levenshtein distance, where segment distances are found on the basis of Barkfilter segment distances, four length gradations are used, diphthongs are represented as a sequence of two segments, and logarithmic segment distances are used.

Examining the recording-based methods we found that the three Levenshtein variants gave less satisfying results. We explained this by the rough way in which word length was normalized and by diversity in voice quality. For the first problem solutions may be found in the field of automatic speech recognition (ASR). The second problem may be solved by using a large number of speakers per variety instead of exactly one speaker, as we did until thus far.

Other future work may consist of carrying out a perception experiment with many more varieties. When validating on the basis of a denser sampling, minor differences may nonetheless emerge clearly.

The best method we found in Chapter 7 was applied to a larger set of 55 Norwegian varieties in Chapter 8. The results were analyzed by clustering and multidimensional scaling. When comparing our results to the authoritative map of Skjekkeland (1997), we found some minor and some major differences. On the one hand, this may be the result of the choice of the 58 words. A better approach would be to select the words randomly from a corpus as Bolognesi and Heeringa (2002) did. On the other hand, the map of Skjekkeland is based on a restricted number of phenomena. We are not sure in how far the map accords with the perception of the speakers. Creating a new Norwegian dialect map on the basis of the arrow method, as done by Daan and Blok (1969) for the Netherlandic part of the Dutch language area, would be interesting.

In Chapter 9 we calculated distances between 360 Dutch variants with the same Levenshtein variant as used for the Norwegian data. Data was taken from the *Reeks Nederlandse Dialectatlassen* (RND). On the basis of these distances cluster analysis was applied and multidimensional scaling was performed. We compared the results to the map of Daan and Blok (1969). We found similarities and differences. Larger groups in our results were divided into smaller groups in the map of Daan, and a larger group in the map of Daan was divided into smaller groups in our results. This suggests that not all borders are equally significant on the map of Daan. When analyzing our results we also examined dendrograms. The benefit of a dendrogram is that groups and borders can be found at any level of significance.

We found it to be a big disadvantage that the RND transcriptions are made by different transcribers. Although we normalized transcriber differences to some extent, we did not succeed eliminating them all, as appeared in our results. We

would like to calculate distances between Dutch varieties again on the basis of data from the *Fonologische Atlas Nederlandse Dialecten* (FAND) (Goossens et al., 1998, 2000). For the compilation of this atlas also different transcribers were involved. Nonetheless, the transcriptions are known to be of excellent quality.

Besides examining the relations between varieties, we also compared the varieties with respect to Standard Dutch. We found results which accord rather well with linguistic reality and with general opinion. The Frisian varieties appeared to be most distant.

From validation work on the one hand, and results from application to Norwegian and Dutch varieties on the other hand, the Levenshtein distance appeared to be a useful tool for finding dialect distances. Differences between our results and existing maps may be explained mostly by shortcomings in our data or in the traditional maps.

10.2 Applications

In this thesis we applied Levenshtein distance to the Norwegian NOS data and to the Dutch RND data. In Bolognesi and Heeringa (2002) Levenshtein distance is applied to a set of 54 Sardinian varieties, Latin, Italian, Genoese, Spanish, Catalan and Dutch. Latin was included since all Romance languages originated from Latin. Italian, Genoese, Spanish and Catalan were included since this languages influenced Sardinian in the past. Dutch was included to show the relative closeness of the Romance languages compared to the Germanic Dutch language. On the basis of Levenshtein distances the varieties were classified. The classification of the Sardinian varieties accorded with dialectological opinion. Since the Sardinian varieties are known to be relatively archaic with respect to the other Romance varieties, they were expected to be very close to Latin. It appeared that Italian was most close to Latin, followed by two Sardinian dialects, Spanish, 39 Sardinian varieties, Catalan, 13 Sardinian varieties, Genoese and – obviously at the end – Dutch. The authors found none of the Sardinian varieties to be obviously more conservative than any of the other Romance varieties in the investigation.

As mentioned above it would be interesting to study the Dutch language area again on the basis the FAND data. However it is a pity that many dialect atlasses or data sets are bounded by political borders, and not by linguistic borders. Inspired by the traditional map in Niebaum and Macha (1999, p. 193), we would like to create a new dialect map of the continental West Germanic language area, including the Netherlands, Flanders, Luxemburg, Germany, Switzerland, Liechtenstein and Austria. Similar to this, it should be interesting to investigate the whole Scandinavian language area, including Iceland, Faroe Islands, Norway, Sweden, Denmark and the Swedish speaking part of Finland. However we would like to enlarge the bounds even more, creating a map of Europe, where each of

the five continua as shown in the map in Chambers and Trudgill (1998, p. 6) and the English continuum are represented in sufficient detail.[2] Possibly the *Atlas Linguarum Europae* (ALE) may be suitable for this purpose (Weijnen et al., 1973–1997).

Besides synchronic measurements, Levenshtein is also useful for diachronic research. In Heeringa and Nerbonne (2000) distances are calculated between 41 Dutch varieties on the basis of old and new transcriptions. The old transcriptions were based on translations of the parable 'The prodigal son' which were compiled by Winkler (1874). The new transcriptions were based on translations of the same text which were compiled by Harrie Scholtmeijer in 1996. Heeringa and Nerbonne used 41 varieties which appeared in both the set of Winkler and in the set of Scholtmeijer. The old and the new varieties were classified. On the basis of the 1874 data a rather sharp division in Frisian, Low Saxon and Low Franconian varieties was found, but for 1996 varieties a division in Frisian, Western Dutch and Eastern Dutch varieties was found. Further an old and a new version of Standard Dutch was added. The old varieties were compared to old Standard Dutch, the new ones to new Standard Dutch. It appeared that the majority of dialects converged to Standard Dutch. Only the dialects along the South-West coast line and in the Middle-East diverged somewhat from Standard Dutch.

In Heeringa et al. (2000) a study about Dutch-German Contact in and around Bentheim is presented. Although the RND mainly contains varieties in the Netherlands and North Belgium, 8 varieties in the German county of Bentheim were also included (see the map in Figure 9.15). The recordings of the varieties in and around Bentheim are made in 1974–1975. In 1999 Heeringa et al. made new recordings of the same Bentheim varieties and 9 varieties at the Dutch side around Bentheim. Standard Dutch and Standard German were added. There were minor differences between the older and the newer version of Standard Dutch, but the older and newer version of Standard German were the same. Just as for the data source mentioned above, the older and newer varieties were classified. The classification results showed that some dialects in the German part, which could be regarded as Dutch Low Saxon dialects in 1974–1975, were found to be German dialects in 1999. On the other hand, Dutch dialects which were grouped among German Low Saxon dialects in 1974–1975, were found to be grouped among the other Dutch dialects in 1999. All Dutch dialects shifted towards Standard Dutch while all Low German dialects shifted towards Standard German. From the results it was concluded that the political border nowadays has got a significant influence on the graduality of the dialect continuum, acting as a separator between Dutch and German dialects.

[2]It is striking that the English continuum including England, Ireland and Scotland is not shown on the map of Chambers and Trudgill (1998, p. 6), although the caption of the figure is: 'European dialect continua'.