Comparison and Classification of Dialects

John Nerbonne and Wilbert Heeringa and Peter Kleiweg
Alfa-informatica, BCN, University of Groningen
9700 AS Groningen, The Netherlands
{nerbonne,heeringa,kleiweg}@let.rug.nl

Abstract

This project measures and classifies language variation. In contrast to earlier dialectology, we seek a comprehensive characterization of (potentially gradual) differences between dialects, rather than a geographic delineation of (discrete) features of individual words or pronunciations. More general characterizations of dialect differences then become available. We measure phonetic (un)relatedness between dialects using Levenshtein distance, and classify by clustering distances but also by analysis through multidimensional scaling.

1 Data and Method

Data is from *Reeks Nederlands(ch)e Dialectatlassen* (Blancquaert and Pée, 1925 1982)). It contains 1,956 Netherlandic and North Belgian transcriptions of 141 sentences. We chose 104 dialects, regularly scattered over the Dutch language area, and 100 words which appear in each dialect text, and which contain all vowels and consonants.

Comparison is based on Levenshtein distance, a sequence-processing algorithm which speech recognition has also used (Kruskal, 1983). It calculates the "cost" of changing one word into another using insertions, deletions and replacements. L-distance (s_1, s_2) is the sum of the costs of the cheapest set of operations changing s_1 to s_2.

sɔəgIrl	delete r	1
sɔəgIl	replace I/ø	2
sɔəgøl	insert r	1
sɔrəgøl		
	Sum distance	4

The example above illustrates Levenstein distance applied to Bostonian and standard American pronunciations of *saw a girl*. Kessler (1995) applied Levenshtein distance to Irish dialects. The example simplifies our procedure for clarity: refinements due to feature sensitivity are omitted. To obtain the results below, costs are refined based on phonetic feature overlap. Replacement costs vary depending on the phones involved. Different feature systems were tested; the results shown are based on Hoppenbrouwers' (SPE-like) features (Hoppenbrouwers and Hoppenbrouwers, 1988).

Comparing two dialects results in a sum of 100 word pair comparisons. Because longer words tend to be separated by more distance than shorter words, the distance of each word pair is normalized by dividing it by the mean lengths of the word pair. This results in a halfmatrix of distances, to which (i) clustering may be applied to CLASSIFY dialects (Aldenderfer and Blashfield, 1984); while (ii) multidimensional scaling may be applied to extract the most significant dimensions (Kruskal and Wish, 1978).

2 Results

We have validated the technique using cross-validation on unseen Dutch dialect data (Nerbonne and Heeringa, 1999). The map in Figure 1 distinguishes Dutch "dialect area" in a way which nonstatistical methods have been unable to do (without resorting to subjective choices of distinguishing features). Ongoing work applies the technique to questions of convergence/divergence of dialects using dialect data from two different periods. Finally, the MDS analysis gives mathematical form to the intuition of dialectologists in Dutch (and other areas) that the material is best viewed as a "continuum". The map is obtained by interpreting MDS dimensions as colors and mixing using inverse distance weighting. Further information on the project is available at www.let.rug.nl/alfa/, "Projects."

3 Acknowledgements

Joseph Kruskal's advice has been invaluable.

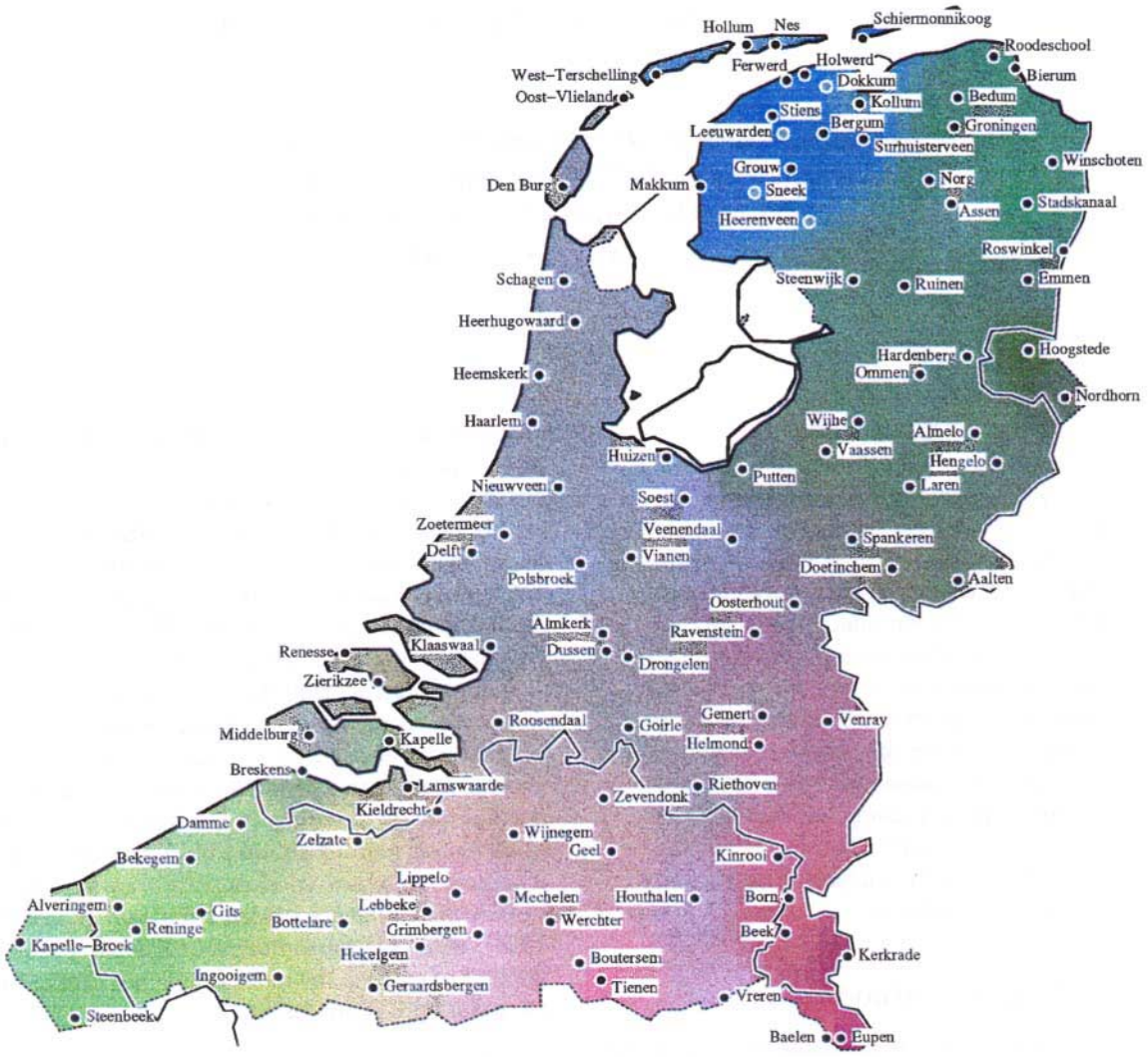

Figure 1: The most significant dimensions in average Levenshtein distance, as identified by multidimensional scaling, are colored red, green and blue. The map gives form to the dialectologist's intuition that dialects exist "on a continuum," within which, however significant differences emerges. The Frisian dialects (blue), Saxon (dark green), Limburg (red), and Flemish (yellow-green) are clearly distinct.

References

Mark S. Aldenderfer and Roger K. Blashfield. 1984. *Cluster Analysis.* Quantitative Applications in the Social Sciences. Sage, Beverly Hills.

E. Blancquaert and W. Pée. 1925-1982. *Reeks Nederlandse Dialectatlassen.* De Sikkel, Antwerpen.

Cor Hoppenbrouwers and Geer Hoppenbrouwers. 1988. De featurefrequentiemethode en de classificatie van nederlandse dialecten. *TABU: Bulletin voor Taalwetenschap*, 18(2):51–92.

Brett Kessler. 1995. Computational dialectology in Irish Gaelic. In *Proc. of the European ACL*, pages 60–67, Dublin.

Joseph Kruskal and Myron Wish. 1978. *Multidimensional Scaling.* Sage, Beverly Hills.

Joseph Kruskal. 1983. An overview of sequence comparison. In David Sankoff and Joseph Kruskal, editors, *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*, pages 1–44. Addison-Wesley, Reading, Mass.

John Nerbonne and Wilbert Heeringa. 1999. Computational comparison and classification of dialects. *Zeitschrift für Dialektologie und Linguistik.* Spec. iss. ed. by Jaap van Marle and Jan Berens w. selections from 2nd Int'l Congress of Dialectologists and Geolinguists, Amsterdam, 1997.