Measuring Dialect Distance Phonetically

John Nerbonne and **Wilbert Heeringa**
Alfa-informatica, BCN
P.O.Box 716, Rijksuniversiteit Groningen
NL-9700 AS Groningen, The Netherlands
nerbonne@let.rug.nl

Abstract

We describe ongoing work in the experimental evaluation of a range of methods for measuring the phonetic distance between the dialectal variants of pronunciations. All are variants of Levenshtein distance, both simple (based on atomic characters) and complex (based on feature vectors). The measurements using feature vectors varied according to whether city-block distance, Euclidean distance or (a measure using) Pearson's correlation coefficient was taken as basic. Variants of these using feature weighting by entropy reduction were systematically compared, as was the representation of diphthongs (as one symbol or two). The results were compared to well-established scholarship in dialectology, yielding a calibration of the method. These results indicate that feature representations are more sensitive, that city-block distance is a good measure of phonetic overlap of feature vectors, that weighting is not useful, and that two-phone representations of diphthongs provide a more satisfactory base for this sort of comparison.

Keywords: dialectology, phonetic (dis)similarity

1 Motivation

Dialectologists frequently speak of the range of dialects they describe as a "continuum", [1] which suggests a need to supersede the inherently discrete method of isoglosses. Dialectologists have long recognized the need for alternative notions of dialectal relationships (Durand (1889),p.49).

[1] For example, Tait on Inuit: "a fairly unbroken chain of dialects [...] the furthest extremes of the continuum being unintelligible to one another" (Tait (1994), p.3)

It is furthermore the case that a sensitive measure of dialectal distance could have broad application to questions in sociolinguistics and historical linguistics, e.g. the significance of political boundaries, the effect of the media, etc.

Levenshtein distance is a measure of string distance that has been applied to problems in speech recognition, bird song ethology, and genetics. It is presented in (Kruskal, 1983), and may be understood as the cost of (the least costly set of) operations mapping from one string to another.

Kessler (1995) applied Levenshtein distance to Irish Gaelic dialects with remarkable success, and Nerbonne et al. (1996) extended the application of his techniques to Dutch dialects, similarly with respectable results. Although Kessler and Nerbonne et al. (1996) experimented with more sensitive measures, their best results were based on calculations of phonetic distance in which phonetic overlap was binary: nonidentical phones contribute to phonetic distance, identical ones do not. Thus the pair [a,t] count as different to the same degree as [a,ɔ].

2 Background

In the interest of space we omit an introduction to Levenshtein distance, referring to (Kruskal, 1983). It may be understood as the cost of (the least costly set of) operations mapping from one string to another. The basic costs are those of (single-phone) insertions and deletions, each of which costs half that of substitutions. Nerbonne et al. (1996) explains its use in the present application at some length. The various modifications below all tinker with the cost of substituting one phone for another.

Kessler (1995) experimented with making the measure more sensitive, but found little progress in using features, for example. The present paper experiments systematically with several variations on the basic Levenshtein theme.

The overall scheme is as follows: a definition of phonetic difference is applied to 101 pairs of words from forty different Dutch dialect areas. All of the pronunciations are taken from the standard dialect atlas ((Blacquart et al, 19251982)—hence: RND, *Reeks Nederlandse Dialectatlassen*). After some normalization, this results in an AVERAGE PHONETIC difference for those dialects—a 40×40 matrix of differences in total (of which one half is redundant due to the symmetry of distance: $\text{dist}(a,b) = \text{dist}(b,a)$). This distance matrix is compared to existing accounts of the dialects in question, especially the most recent systematic account, (Daan and Blok, 1969). A visualization tool normally identifies very deviant results, see Fig. 1. Finally the distance matrix is subjected to a heuristic clustering algorithm as a further indication of quality.[2]

3 Refinements for Dialectology

The dialects are compared on the basis of the words of 101 items. So the total distance of two dialects is equal to the sum of 101 Levenshtein-distances. If we simply use the Levenshtein-distance, it would tend to bias measurements so that changes in longer words would tend to contribute more toward the average phonetic distance (since they tend to involve more changes). This may be legitimate, but since words are a crucial linguistic unit we chose to stick to average word distance. This involves the computation of 'relative distance', which we get by dividing the absolute distance by the length of the larger word. We have also considered using the average length of the two words being compared, which makes little difference where both words are present.

Missing words pose a problem as does lexical replacement. We wished to handle these consistently (to obtain a consistent measure of distance), even recognizing the danger of conflating phonetic and lexical effects. Throughout this paper we do conflate the two, reasoning that this is the lesser of two evils—the other of which is deciding when massive phonetic modification amounts to lexical difference.

Naturally no difference is recorded where a word is missing in both dialects. If only one dialect is missing the word, the difference at that point is just *length* $\times$ insertion-cost, but normalization divides this by the *length* again, yielding just the cost of insertion. This is a point at which the decision noted above–to obtain relative distance via Levenshtein distance divided by longer length–is important. Recall the alternative mentioned there, that of relativizing to the average length. This would double the distance measured in cases where words are missing, biasing the overall distance toward dialects with less lexical overlap. This seemed excessive.

Similarly, for some items there are two words possible. If dialect 1 has word1a and word1b, and dialect 2 has word2, we calculate the distance by averaging *distance*(word1a,word2) and *distance*(word1b,word2). If both dialect 1 and dialect 2 have multiple variants, we average all pairs of distances.

Although we experimented with variable costs for substitutions, depending on whether their base segments or diacritics differ, we could not settle on a natural weighting, and further reasoned that a feature-based cost-differential should systematize what the transcription-based differential intended. This is resumed below.

Dutch has a rich system of diphthongs, which, moreover have been argued to be phonologically disegmental (Moulton, 1962). We therefore experimented both with single-phone and two-phone diphthongal representations. It turned out the representations with two phones were superior (for the purposes of showing dialectal relatedness).[3]

3.1 Feature Vectors

If we compare dialects on the basis of phonetic symbols, it is not possible to take into account the affinity between sounds that are not equal, but are still related. Methods based on phonetic symbols do not show that 'pater' and 'vader' are more kindred then 'pater' and 'maler'. This problem can be solved by replacing each phonetic symbol by a vector of features. Each feature can be regarded as a phonetic property which can be used for classifying of sounds. A feature vector contains for each feature a value which indicates to what extent the property is instantiated. Since diacritics influence feature values, they likewise figure in the mapping from transcriptions to feature vectors, and thus automatically figure in calculations of phonetic distance.

In our experiment, we have used the feature vectors which are developed by (Vieregge, A.C.M.Rietveld, and Jansen, 1984) (we earlier used the SPE features as these were modified for dialec-

[2]The choice of clustering technique is important, but is not the focus of the present paper. The methods here were compared using Ward's method, a variant of hierarchical agglommerative clustering which minimizes squared error. See (Jain and Dubes, 1988) for clustering techniques.

[3]It would be rash to argue from this to any phonological conclusion about the diphthongs. The two-phone representation makes it easier to measure related pronunciation, and this is probably why it suits present purposes better.

tology use by (Hoppenbrouwers and Hoppenbrouwers, 1988), but obtained distinctly poorer results in spite of the larger number of features). Vieregge et al. make use of 14 features [longer discussion of Vieregge's system as well as the translation transcriptions in the RND in full version of paper].

We compare three methods for measuring phonetic distance. The first is MANHATTAN DISTANCE (also called "taxicab distance" or "city block" distance). This is simply the sum of all feature value differences for each of the 14 features in the vector. $\delta(X,Y) = \sum_{i=1}^{n} |X_i - Y_i|$

Second, we tried EUCLIDEAN DISTANCE. As usual, this is the square root of the sum of squared differences in feature values. $\delta(X,Y) = \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2}$

Third, we examined the Pearson correlation coefficient, r. To interpret this as distance we used $1-r$, where r is the usual $\frac{1}{n-1}\sum(\frac{x-\bar{x}}{s_x})(\frac{y-\bar{y}}{s_y})$.

In the Levenshtein algorithm based on symbols, three operations were used: 'substitution', 'insertion' and 'deletion'. A substitution was regarded as a combination of an insertion and a deletion, so substitutions counted two, and "indels" one. When we compare vectors instead of phonetic symbols, the value for a substitution is no longer a fixed value, but varies between two extremes. However, for indels we have to choose a fixed value as well. This value was estimated by calculating the average of the values of all substitutions which take place in the comparison proces, and dividing this average by 2.

3.2 Information-Gain Weighting

Not all features are equally important in classifying the sounds used in the dialects. For example, it turned out that no positive value for the feature [*flap* ±] occurred in any of the words in the dialects examined. We therefore experimented with weighing each feature by information gain, a number expressing the average entropy reduction a feature represents when known (Quinlan, 1993; Daelemans et al., 1996).

To calculate this we need a base figure for database entropy:

$$H(D) = -\sum_{i} p_i log_2 p_i$$

If we have n different vectors for all the dialects, then $1 \leq i \leq n$. p_i is the probability of vector i, estimated by its frequency divided by $|D|$, which is the total number of vectors in all dialects.

Second we calculate the average entropy for each feature:

$$H(D_{[f]}) = \sum_{v_i \in V} H(D_{[f=v_i]}) \frac{|D[f=v_i]}{|D|}$$

$|D_{[f=v_i]}|$ is the number of vectors that have value v_i for feature f. V is the set of possible values for feature f. $H(D_{[f=v_i]})$ is the remaining entropy of all vectors in the database that have value v_i for feature f. It is calculated using the first formula, where the i's are now only the vectors that have value v_i for feature f.

Finally we can calculate the information gain associated with a feature:

$$G(f) = H(D) - H(D_{[f]})$$

If we then compare two vectors using Manhattan distance, the weighted difference between two vectors X and Y is now:

$$\Delta(X,Y) = \sum_{i=1}^{n} G(f_i)|X_i - Y_i|$$

And similarly for Euclidean distance and "inverse correlation".

We have recently become aware of the work of Broe (1996), which criticizes the simple application of entropy measures to feature systems in which some features are only partially defined. Such phonological features clearly exist: e.g., [`lateral`] and [`strident`] apply only to consonants and not to vowels. Broe furthermore develops a generalization of entropy sensitive to these cases. This is an area of current interest.

4 Experiments

The dialect varieties were chosen to contain "easy" cases as well as difficult ones. Frisian is accepted as rather more distinct from other areas, and eight Frisian varieties are represented in the wish to see quickly that that distance metrics could distinguish these. The full list of variants may be seen in Fig. 1.

5 Results

We compared a total of 14 methods, shown in Table 1. While none of these performed very poorly, several tendencies emerge.

- Two-phone representations of diphthongs outperform single-phone representations
- Unweighted representations outperform representations to which weightings were added. This is surprising.
- Manhattan distance narrowly outperforms "correlation" which narrowly outperforms Euclidean distance.

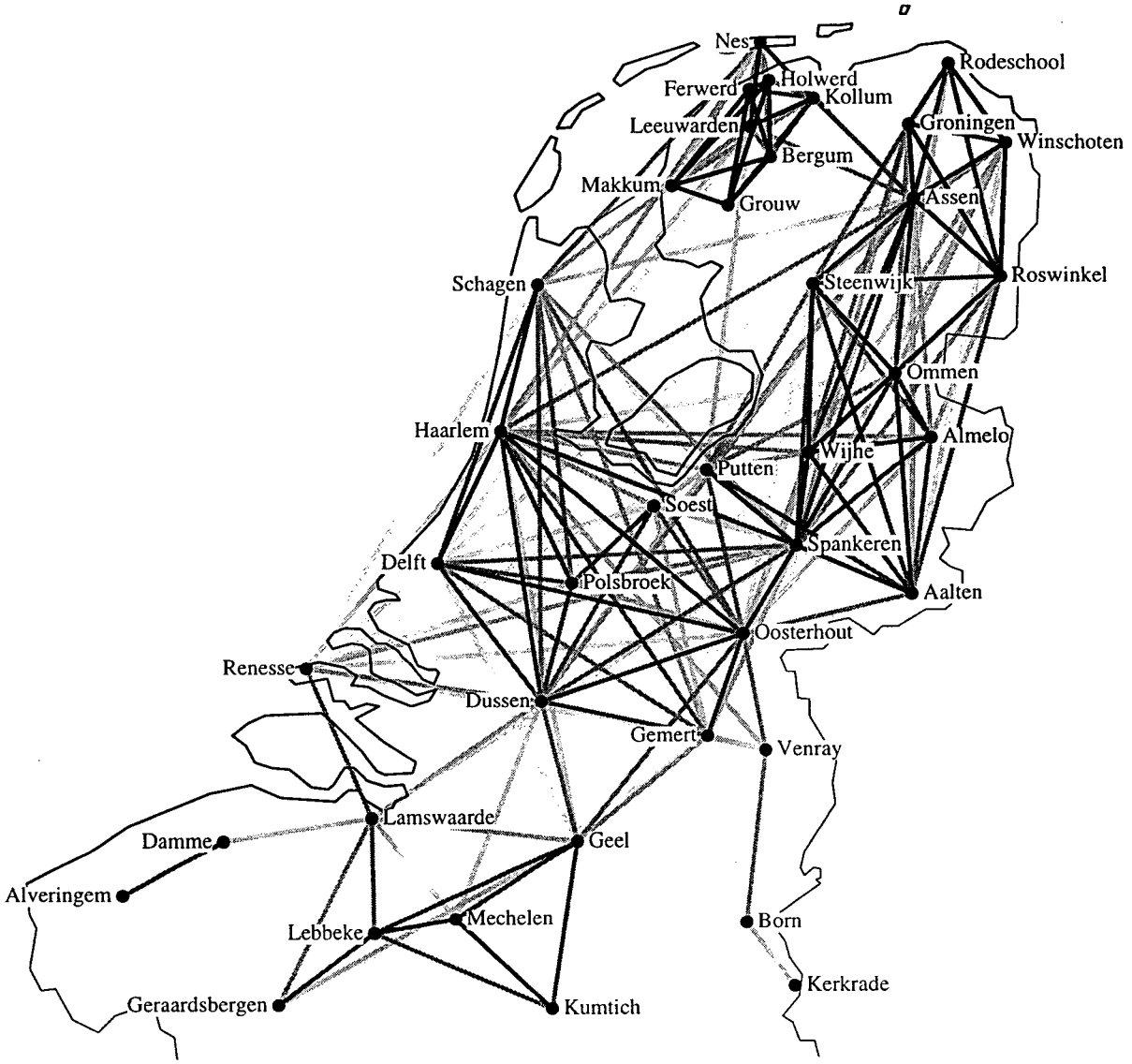

Figure 1: The distances calculated between the 40 pairs of Dutch dialects. Darker lines correspond to phonetically proximate language variants, lighter lines to more distant variants. Phonetically very distant relations result in lines too faint to be seen (in the interest of overall contrast). The Frisian dialects (in the Northwest) are clearly recognized as relatively close, as are the Saxon dialects (Northeast). The situation in the South is less simple.

	phone/feature-based	weighted	feature-comparison	diphthong
1	phones			one phone
2	phones			two phones
3	features	no	Manhattan	two phones
4	features	no	Manhattan	one phone
5	features	no	Euclidean	two phones
6	features	no	Euclidean	one phone
7	features	no	correlation	two phones
8	features	no	correlation	one phone
9	features	yes	Manhattan	two phones
10	features	yes	Manhattan	one phone
11	features	yes	Euclidean	two phones
12	features	yes	Euclidean	one phone
13	features	yes	correlation	two phones
14	features	yes	correlation	one phone

Table 1: Fourteen variants of Levenshtein distance which were compared in the task of distinguishing Dutch dialect distances. Top performer (3) used features in place of discrete segments, no information-gain weighting, Manhattan (city-block) distance, and a two-segment representation of diphthongs.

Thus, method (3) was best.

The superiority is seen in the degree to which the distance matrices and resulting dendrograms match those of expert dialectologists, in particular, (Daan and Blok, 1969).[4]

We did not apply a measure to the degree of coincidence between the experts' division into dialect groups and the grouping induced by the Levenshtein distance metric. Instead, we compared the dendrogram to the dialect map and checked for congruence. Some of the results accord better with expert opinion.

For example, dialectologists generally locate Delft as closer to Haarlem and Schagen (than to Oosterhout, Dussen and Gemert). The better distance measures do this as well, but not several of the weighted measures. The weighted measures and the unweighted correlation-based measures similarly failed to recognize the coastal (western) Flemish subgroup (*Westflaams* or *Zeeuwsvlaams*), represented in our data set by Alveringem, Damme, Lamswaarde, and Renesse.

Daan's work is accompanied by a map that also appears in the *Atlas of the Netherlands*, as *Plate X-2*.[5] It divides the Dutch area into 28 areas of roughly comparable dialect regions. Furthermore, it uses colortones to denote relative distance from standard Dutch. This information can be used to further calibrate the methods here. First, the relative distance from standard Dutch (given in colortones) can be translated to predictions about relative phonetic distance. For example, *Twents* is shaded dark green (and is represented in our data set by the the dialect spoken in Almelo), while *Veluws* is shaded light green (and is represented by Soest and Putten). There is an intermediate dialect, *Gelders-Overijssels* shaded an intermediate green and represented by Ommen, Wijhe and Spankeren. These relative distances (to ABN, represented in our data set by Haarlem and Delft) should be reflected in Levenshtein distance, and we can test the prediction by how accurate the relfection is. This method of testing has the large advantage that it tests only Levenshtein distance without involving the added level of clustering.

A second method of using the dialect map to calibrate the Levenshtein metric is to use the 28 various dialect regions as predictions of "minimal distance". Here we can compare the map most simply to the dendrogram. In the present work, it may be noted that the Frisian dialects and the dialect of Groningen-North Drenth are indeed identified as

[4] It should be noted that Daan and Blok (1969) incorporate native speakers' subjective judgements of dialect distance in their assessment (their "arrow method"). But their final partition of dialects into differenct groups is well-accepted.

[5] Printed by the *Topografische Dienst*, Delft, 1968.

groups (by the Levenshtein method combined minimal error clustering). It is more difficult to use the dialect map in this way without using the dendrogram as well. In particular, it is not clear how the borders on the dialect map are to be interpreted. Keeping in mind the "continuum" metaphor noted in Sec. 1, the borders cannot be interpreted to be marking partitions of minimal distance. That is, it will not be the case that each pair of elements in a given cluster are closer to each other than to any elements outside.

An interesting fact is that while no very close correlation is expected between dialectal distance and geographical distance, still the better techniques generally correlated higher with geographic distance than did the poorer techniques (at approx. $r = 0.72$).

We conclude that the present methods perform well, and we discuss opportunities for more definitive testing and further development in the following section.

6 Future Directions

We should like to extend this work in several directions.

- We should like to find a way to measure the success of a given distance metric. This should reflect the degree to which it coincides with expert opinion (which is necessarily rougher). See Sec. 5.
- An examination of grouping methods is desirable.
- The present method averages 101 word distances to arrive at a notion of dialect difference. It would be interesting to experiment directly with the 101-dimensional vector, standardized to reflect the distance to standard Dutch (*algemeen beschaafd Nederlands*, ABN) and using, e.g., the $\cos(\vec{x}, \vec{y})$ as a distance measure (on vectors whose individual cells represent Levenshtein distances from ABN pronunciations).
- For more definitive results, the method should be tested on material for which it has NOT been calibrated, ideally a large database of dialectal material.
- Finally, it would be interesting to apply the technique to problems involving the influence of external factors on language variation, such as migration, change in political boundaries, or cultural innovation.

7 Acknowledgements

We thank Peter Kleiweg for his graphic programs (seen in all of the figures here), and thanks also to an anonymous reviewer for comments.

References

Blacquart et al, E. 1925/1982. Reeks nederlandse dialectatlassen.

Broe, Michael. 1996. A generalized information-theoretic measure for systems of phonological classification and recognition. In *Computational Phonology in Speech Technology: Proceedings of the Second Meeting of the ACL Special Interest Group in Computational Phonology*, pages 17–24, Santa Cruz. Association for Computational Linguistics.

Daan, Jo and D. P. Blok. 1969. *Van Randstad tot Landrand; toelichting bij de kaart: Dialecten en Naamkunde.* Amsterdam: Noord-Hollandsche Uitgevers.

Daelemans, Walter, Jakub Zavrel, Peter Berck, and Steven Gillis. 1996. Memory-based part of speech tagging. In Gert Durieux, Walter Daelemans, and Steven Gillis, editors, *Proc. of CLIN '95.* Antwerpen, pages 185–202.

Durand, Jean-Paul. 1889. Notes de philologie rouergate, 18. *Revues des Langues Romanes*, 33:47–84. cited by Kessler.

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

Jain, K. and R. C. Dubes. 1988. *Algorithms for clustering Data.* Englewood Cliffs, New Jersey: Prentice Hall.

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

Kruskal, Joseph. 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.* Addison-Wesley, Reading, Mass., pages 1–44.

Moulton, William. 1962. The vowels of dutch: Phonetic and distributional classes. *Lingua*, 11:294–312.

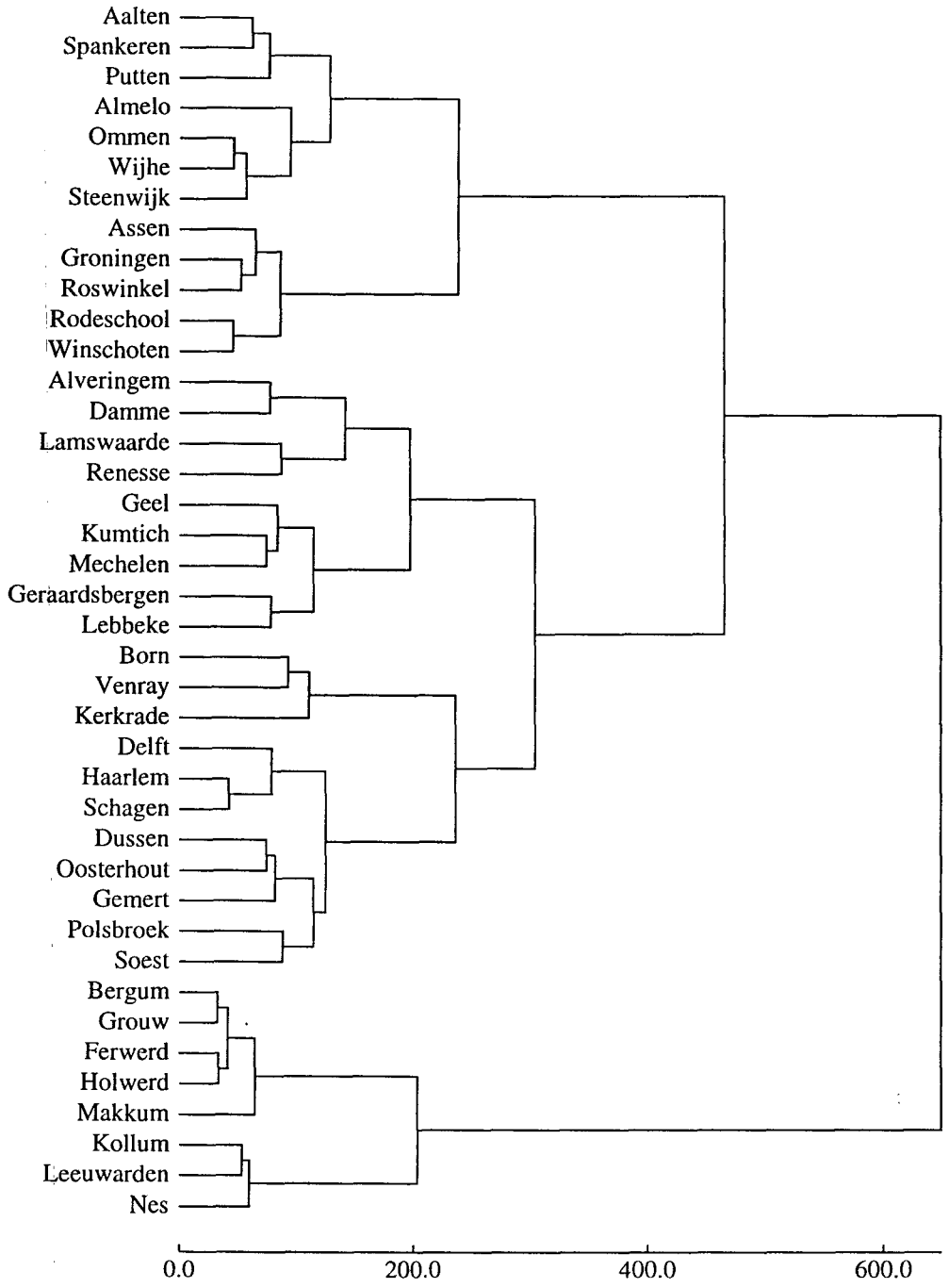

Figure 2: A dendrogram derived from the distance matrix based on (unweighted) Manhattan distance between feature representations. Note that Frisian and Dutch variants are distinguished most significantly, while within Dutch there major distinctions are Lower Saxon dialects (top), Flemish, and Franconian (lowest of the three most significant branches within Dutch). This accords well with dialectal scholarship. The dendrogram was obtained using a Ward's method of hierarchical agglommerative clustering, a minimized square-error method. Alternative clustering methods have also been compared in this project, but that topic is beyond the bounds of this paper.

Figure 3: The four most significant dialect groups isolated bu this method correspond to Frisian (northwest, dark), Lower Saxon (northeast,light), Franconian (central, light-intermediate) and Flemish (south, dark-intermediate).

Nerbonne, John, Wilbert Heeringa, Erik van den Hout,
Peter van der Kooi, Simone Otten, and Willem van de Vis. 1996. Phonetic distance between dutch dialects. In Gert Durieux, Walter Daelemans, and Steven Gillis, editors, *Proc. of CLIN '95*. Antwerpen, pages 185–202. Also available as `http://grid.let.rug.nl/~nerbonne/papers/dialects.ps`.

Quinlan, J. Ross. 1993. *C4.5: Programs for Machine Learning*. San Mateo: Morgan Kaufmann.

Tait, Mary. 1994. North America. In Christopher Moseley and R.E. Asher, editors, *Atlas of the World's Languages*. Routledge, London and New York, pages 3–30.

Vieregge, Wilhelm H., A.C.M.Rietveld, and Carel Jansen. 1984. A distinctive feature based system for the evaluation of segmental transcription in dutch. In *Proc. of the 10th International Congress of Phonetic Sciences*, pages 654–659, Dordrecht.