INTRODUCTION TO REISSUE EDITION

EDIT DISTANCE AND DIALECT PROXIMITY

John Nerbonne
with Wilbert Heeringa and Peter Kleiweg

David Sankoff and Joseph Kruskal's *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison* (hereafter *Time Warps*) is a young (1983) classic which has inspired developments in computer science, pure and applied linguistics, computational biology, and even music and ethnology. CSLI Publications deserves the appreciation of all these scientific subfields for undertaking its republication. Because its first chapter, "Overview," is Joseph Kruskal's gentle introduction to Levenshtein distance (also known as sequence distance and edit distance) (Kruskal 1999), the book will be useful to students as well as researchers. The overview explores the concept of sequence distance first from the perspective of alignment – finding the correspondence between sequences that minimizes distance. It goes on to introduce a dynamic programming algorithm which efficiently calculates the distance, and provide notes on the history of this and related concepts. Later chapters in *Time Warps* go on to explore several interesting extensions and applications in depth, including recognizing relatedness in DNA, matching speech input to lexical hypotheses, tracking the development of bird songs over time, and to error corrections (e.g., of keyboard input). Some of these chapters introduce significant refinements in order to handle transposed elements, continuous input (rather than sequences of discrete elements), and tree-structure.

This introduction to the new edition of *Time Warps* in *The David Hume Series of Philosophy and Cognitive Science Reissues* focuses on linguistic applications. We shall mention some of the developments since 1983 and suggest why the theory of sequence distance still enjoys enormous popularity in the diverse linguistic fields in which it is applied. The examples come from pure and applied linguistics, including computational linguistics, and draw on work done in Groningen. We shall unfortunately ignore the extensive and very successful work on sequence distance in

computational biology to explore which DNA sequences are likely mutations of one another (Waterman 1989, Waterman 1995, Crochemore and Gusfield 1994, Farach-Colton 1998.)

1 COMPUTATIONAL LINGUISTICS

Levenshtein distance and various derivatives give the researcher a tool with which to measure the relative similarity of various sequences. The technique is general enough to apply to sequences of all sorts.

Pronunciations consist of sequences of sounds, so it is not surprising that *Time Warps* includes Kruskal and Liberman's explication of dynamic time warping, the application of Levenshtein distance to speech recognition (Kruskal and Liberman 1999). Although Hidden Markov Models have largely replaced dynamic time warping in speech recognition, Levenshtein distance is still used in speech recognition, e.g., in order to score the relative closeness of hypotheses to (annotated) correct answers (Veldhuijzen van Zanten et al. 1999). Without this, it is difficult to distinguish recognition results beyond the level of 'correct vs. incorrect'. Other researchers have used variants of Levenshtein distance to diagnose potentially pathological pronunciation deviation (Connolly 1997). The interest in speech pathology requires that deviant pronunciations be aligned with normal ones so that researchers and therapists can explore the specific "operations" that appear to be responsible for differences.

The last ten years have seen an explosion in the availability of large text corpora (Church and Mercer 1993, Klavans 1996, Nerbonne 1998). Among these, there has been special interest in parallel, bilingual corpora (Veronis to appear, 2000). Careful analysis of parallel corpora yields information such as nearest translation equivalents, common usage of words and phrases, and subtleties of grammatical use, all of which is of course useful to translators and developers of translation support software, to bilingual dictionary compilers, and to developers of related applications such as computer-assisted language learning. Veronis (2000) has papers on all of these applications. The fundamental technical problem in supporting the extraction of such information is the problem of aligning texts so that users can easily identify corresponding parts, and the most successful approaches have been derived from the dynamic programming algorithm for calculating edit distance (Gale and Church 1993). Gale and Church's (1993) algorithm take sentences as fundamental units, and proceed from the assumption that sentence length in characters will correspond probabilistically in parallel texts. Special accommodation is made for the sentences' expanding, contracting and merging in translation, but the authors are explicit about the technique's legacy from *Time Warps*. The topic of alignment is important enough so that alternative methods are still actively sought (Church 1993, Chen 1993), and new applications in multimedia are under investigation in which the alignment of text plays an important role (Braschler and Schauble 1998).[1]

[1]The EU OLIVE project `http://twentyone.tpd.tno.nl/olive/` tries to align notes,

2 DIALECTOLOGY

Linguistics studies language variation systematically as it correlates with any number of extra-linguistic variables, but especially with geography and social variables such as class, age, gender, social network, education and trade. The oldest of these is dialectology, the study of how language varies geographically. Dialectology is not only the most senior of the linguistic variation fields, it is one of the oldest branches of linguistics – the early works date from the 19th century as do many of the techniques. See Petyt (1980) and Niebaum (1983).

Traditional dialectology has focused on recording the variety of speech forms within a given area. This has intrinsic, linguistic interest, but it is also interesting because language variety reflects cultural influence, most obviously in the case of migrations, conquests and the creation of states and borders. Dialectology has amassed large amounts of data concerning the specific geographical distribution of individual words, word forms, syntactic constructions, pronunciations and the like. These are normally present in the form of maps showing which forms are used in which areas. In case the same areas spring out repeatedly, one may speak of a "dialect area."

Unfortunately, detailed maps showing the distribution of words and sounds normally fail to distinguish areas, which leads to a scientific puzzle: linguists and lay people alike feel that they recognize varieties, such as Bavarian German, or the Amsterdam variety of Dutch, or the coastal New England variety of English. But these are not the areas consistently picked out by varying forms. There are usually patterns of variation that run criss-cross, ignoring the well-established boundaries of areas.

This puzzle was, as it turned out, a scientific disappointment. Dialectology was originally pursued *inter alia* in order to test the Neo-grammarian hypothesis that sound change is regular. Among its earliest results is that variation is often geographically irregular, i.e., the areas in which different forms are used often turn out to be (perhaps slightly) different areas (Bloomfield 1933, p.322).

Naturally, some dialectologists responded by collecting more data, hoping to find the right distinguishing features. This has been done to such extremes that even friendly reviewers have warned of the dangers of "atomization" (Coseriu [1]1956, 1975, p.50). Lacking analytical techniques with which distributions of different forms can be compared or combined, the result is that hundreds of forms give hundreds of perspectives – in none of which aggregates such as Bavarian German, Amsterdam Dutch, or coastal New England are recognizable combinations of the atoms.

In many cases, however, it often turns out that changes are cumulative, i.e., that more geographically remote areas are linguistically less similar as well so that it is reasonable to speak of a dialect "continuum" (Chambers and Trudgill 1980, § 1.3,§ 8.1-8.6). But this notion – like other aggregate notions – has resisted satisfactory theoretical formulation.

---

speech recognition results and closed captions for the hearing impaired in order to align associated video material (Netter 1998).

This very brief sketch outlines two theoretical questions in dialectology, how aggregate notions such as the variety of a particular area may be approached, and how those aggregates might be seen to exist on a continuum of linguistic variety. We cannot offer theoretical reconstructions of either of these notions, but, with the help of Levenshtein distance, we can measure differences associated with them.

2.1 Measuring dialect difference

When you can measure, [. . .] you know something—Lord Kelvin

Levenshtein distance is employed to obtain a measure of distance in pronunciation. In contrast to the "same/different" classification of earlier dialectology, the measurements are numerical and may therefore be summed and/or averaged to obtain an aggregate characterization of differences between entire varieties (based on larger sets rather than one word at a time). We obtain then a measure of difference rather than a geographic delineation of (discrete) features of individual words or pronunciations. More general characterizations of dialect differences then become available.

A mathematical perspective may be useful. Most linguistic data is naturally nominal, divided into different categories, which allow no natural aggregation. But once the data can be approached numerically – as the use of Levenshtein distance allows, then addition of atomic differences can be interpreted, so that we can sensibly speak of the distance between entire varieties. The measurement also allows us to approach the question of cumulativity more exactly. If dialectologists have been correct in speaking of cumulative differences, then we should likewise find Levenshtein difference increasing with geographical distance – a reflection of the "continuum."

2.2 Data and method

The *Reeks Nederlands(ch)e Dialectatlassen* (Blancquaert and Pée 1925–1982) 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.

As the introductory chapter of this book best explains, Levenshtein distance in its most basic variant involves calculating the "cost" of changing one word into another using insertions, deletions and replacements. Levenshtein distance (s_1, s_2) is the sum of the costs of the cheapest set of operations changing s_1 to s_2. The example below illustrates Levenshtein distance applied to Bostonian and standard American pronunciations of *saw a girl*. Boston pronunciation inserts an [r] between *saw* and *a*, deletes the postvocalic [r] in *girl*, and replaces the short vowel in *girl* with a fronted rounded vowel [ø], like the first vowels in German *Köln* or French *meuble*.

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

Kessler (1995) applied Levenshtein distance to Irish dialects.

The example above simplifies the procedures actually used for clarity: the actual measurements are sensitive to the phonetics of the pronunciations of the basic sounds (*t,d,i,* etc.). To obtain a more sensitive measure, costs are refined based on phonetic feature overlap, so that, e.g. replacing a [d] with a [t] is less costly than replacing the [d] with, say [i]. Replacement costs thus vary depending on the basic sounds involved.

This is done in a way that is standard in linguistics. Each sound is represented by a vector of values for phonetic features. The table below suggests how this system gives rise to characterizations of the overall difference between segments: [i] and [e] are much closer than [i] and [u].

	i	e	u	i-e	i-u
advancement	2(front)	2(front)	6(back)	0	4
high	4(high)	3(mid)	4(high)	1	0
long	3(short)	3(short)	3(short)	0	0
lip-rounding	0(none)	0(none)	1(rounded)	0	1

To avoid making the measurements too dependent on one feature system as opposed to another, two feature systems were tested; the results analyzed below are based on Hoppenbrouwers' (SPE-like) features (Hoppenbrouwers and Hoppenbrouwers 1988), but a feature system developed by Vieregge to measure transcriber accuracy also yields very good results (Vieregge et al. 1984). The distance between two vectors of feature values was taken to be the Euclidean distance between the two, $\delta(X, Y)$, where:

$$\delta(X, Y) = \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2}$$

Several further refinements were the subject of experimentation, including the optimal representation of diphthongs (one segment vs. two), the value of weighting by frequency and/or information gain, and the use of the standard language as a calibration in measurements (Nerbonne and Heeringa to appear, 1999b). We have begun validating the technique, including these aspects, using cross-validation on unseen Dutch dialect data (Nerbonne and Heeringa 1999a).

2.3 Results

Comparing two varieties 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 length of the word pair. This results in a half-matrix of distances, which is best visualized in Figure 1. This fulfills the first desideratum above, obtaining a means of aggregating over the atomistic differences between individuals words. Using the Levenshtein measure, we can now say something about the overall relation between two varieties – even when individual elements may vary and even be inconsistent in the tendencies they display.

A further goal was to explore how dialect areas might be viewed. The half-matrix of distances above was therefore subjected to hierarchical clustering, using Ward's method, which minimizes the error introduces by clustering (Aldenderfer and Blashfield 1984). This classifies the dialects into relatively close groups. The most significant groups which emerge from clustering correspond to those identified in traditional dialectology, viz., Frisian, Low Saxon, and Franconian (Nerbonne and Heeringa to appear, 1999b). See Figure 2. We take this to validate the techniques, but current work seeks more direct validation.

Finally, we explore the degree to which the techniques confirm the view of dialects as existing on a "continuum." The most direct means of testing this is to examine whether phonetic distance increases with geographical distance, to see whether distances are indeed "cumulative." Pearson's product-moment correlation coefficient shows that geographic and phonetic distance correlate at a level of $r = 0.68$. Given the large amount of data, it is not surprising that this is a statistically significant level. But it is also a fairly strong correlation, accounting for nearly 45% of the variance. The heights of children and parents correlate to about this degree. So we obtain an answer: the "continuum" is justified in its generalization that phonetic differences are cumulative. These results suggests that one may indeed combine the detailed dialectology maps, showing discrete dialect areas with the often expressed view that dialects exists on a continuum, in accord with Chambers and Trudgill's (1980:127) (sceptical) wishes. These are then compatible ways of viewing the dialectal facts.

To explore further the view of dialects as a continuum, we exploit techniques Joseph Kruskal developed in multidimensional statistical analysis, viz., multidimensional scaling (MDS). MDS may be applied to distance matrices such as the matrix of average Levenshtein distances in order to identify a small set of most significant dimensions (Kruskal and Wish 1978). The map shown in this book's frontispiece (see inside cover) distinguishes Dutch "dialect areas" in a way which non-numerical methods have been unable to do (without resorting to subjective choices of distinguishing features). 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. Since this assigns colors only to the sites for which pronunciations are available, we interpolate over other areas using inverse distance weighting.

Figure 1: The average Levenshtein distance between Dutch dialects. The darker the line between two varieties, the closer they are in (phonetic) Levenshtein distance. Frisian (top center) emerges clearly as a relatively distinct, but internally coherent group.

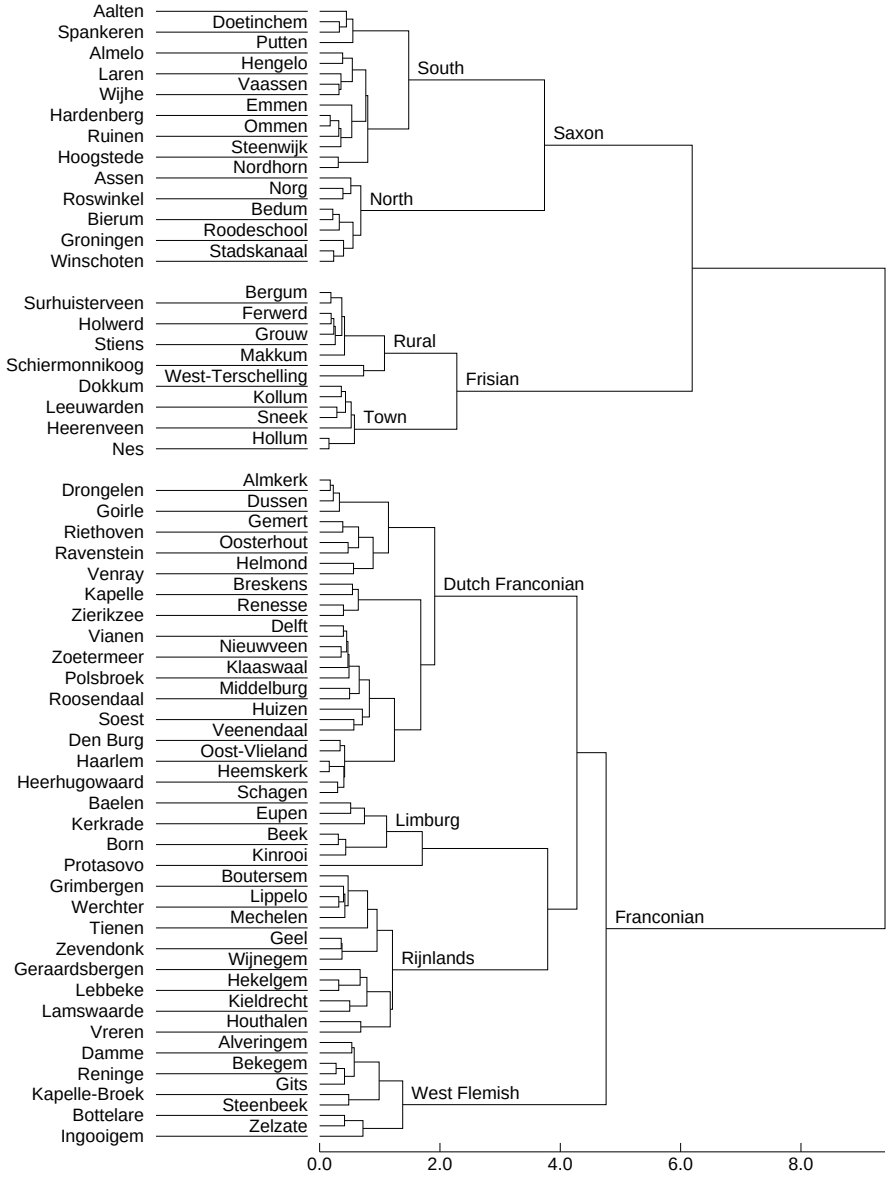


Figure 2: The result of clustering the matrix of Levenshtein distances. The traditional dialect areas (labeled) emerge distinctly. Within the Limburg cluster one finds Protasovo, an emigrant Low German variety, the language of Siberian Mennonites investigated by Nieuweboer (1998) (included out of curiosity).

2.4 Conclusions and prospects in dialectology

We have studied the consistency of the measurements, which, given a set of 100 words, is reliable. In order to validate techniques more rigorously, we have run the techniques on unseen data, and we are currently seeking an objective measure of quality to distinguish the various refinements of Levenshtein distance discussed above. To do this, we shall take measures on a set of variants whose classification into dialect area (Frisian, Saxon, Flemish or Franconian) is uncontested. The question is the degree to which the Levenshtein measure will jibe with the classification. We shall report on this work when it is completed.

Ongoing work applies the technique to questions of convergence/divergence of dialects using dialect data from two different periods. In order to be more generally useful, the distance measure should be applied to the data of other languages. We are also interested in exploring the relation between this (linguistically objective) measure of distance with psycho-acoustical investigations into the perception of varieties (Gooskens 1997).

Further information on the project, also material (data) and some software, is available at `http://www.let.rug.nl/alfa/`, "Projects."

3 *TIME WARPS*

Levenshtein distance is still being applied to new areas, more than 15 years after *Time Warps*'s original publication. We may hope that the continued availability of Sankoff and Kruskal's seminal volume will spur further refinement and application of sequence distance measures.

ACKNOWLEDGEMENTS

Wilbert Heeringa and Peter Kleiweg have implemented and maintained all of the dialectology programs described here. Dicky Gilbers, Tjeerd de Graaf, Jack Hoeksema, Wouter Jansen, Brett Kessler, Joseph Kruskal, Hermann Niebaum and Harry Scholtmeier have offered valuable criticism and advice at various points in this project.

REFERENCES

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

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

Bloomfield, Leonard. 1933. *Language*. New York: Holt, Rhinehart and Winston.

Braschler, M. and P. Schauble. 1998. Multilingual Information Retrieval Based on Document Alignment Techniques. *Proc., 2nd European Conference, (ECDL)* 183–199.

Chambers, Jack, and Peter Trudgill. 1980. *Dialectology*. Cambridge: Cambridge University Press.

Chen, Stanley. 1993. Aligning Sentences in Bilingual Corpora using Lexical Information. *Proc. of the 31st ACL* 9–17. Columbus, OH: ACL

Church, Kenneth. 1993. char_align: A Program for Aligning Parallel Texts at the Character Level. *Proc. of the 31st ACL* 1–8. Columbus, OH: ACL

Church, Kenneth and Robert Mercer. 1993. Introduction to the Special Issue on Computational Linguistics using Large Corpora. *Computational Linguistics* 19(1):1–24. (See articles in this and in following number 19(2))

Connolly, John H. 1997. Quantifying Target-Realization Differences. *Clinical Linguistics and Phoentics* 11:267–298

Coseriu, Eugenio. [1]1956, 1975. *Die Sprachgeographie*. Tübingen: Gunter Narr.

Crochemore, Maxime, and Dan Gusfield (ed.). 1994. *Combinatorial Pattern Matching: the 5th Annual Symposium*. Berlin and Heidelberg: Springer.

Farach-Colton, Martin (ed.). 1998. *Combinatorial Pattern Matching: 9th Annual Symposium*. Berlin and Heidelberg: Springer.

Gale, William A., and Kenneth W. Church. 1993. A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics* 19(1):75–102.

Gooskens, Charlotte. 1997. *On the Role of Prosodic and Verbal Information in the Perception of Dutch and English Language Varieties*. Doctoral dissertation, Nijmegen University.

Hoppenbrouwers, Cor, and Geer Hoppenbrouwers. 1988. De Featurefrequentiemethode en de Classificatie van Nederlandse Dialecten. *TABU: Bulletin voor Taalwetenschap* 18(2):51–92.

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

Klavans, Judith and Philip Resnick, eds. 1996. *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*. Cambridge, MA: MIT Press.

Kruskal, Joseph. [1983] 1999. An Overview of Sequence Comparison. In *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*, ed. David Sankoff and Joseph Kruskal. 1–44. Reprint, with a foreword by John Nerbonne, Stanford, CA: CSLI Publications.

Kruskal, Joseph, and Mark Liberman. [1983] 1999. The Symmetric Time-Warping Problem: From Continuous to Discrete. In *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*, ed. David Sankoff and Joseph Kruskal. 125–161. Reprint, with a foreword by John Nerbonne, Stanford, CA: CSLI Publications.

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

Nerbonne, John, ed. 1998. *Linguistic Databases*. Stanford: CSLI Publications.

Nerbonne, John, and Wilbert Heeringa. 1999a. 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.

Nerbonne, John, and Wilbert Heeringa. to appear, 1999b. Computationele Vergelijking en Classificatie van Dialecten. *Taal en Tongval* 51.

Netter, Klaus. 1998. POP-EYE and OLIVE: Human Language as the Medium for Cross-lingual Multimedia Information Retrieval. *Proc. of 2nd International Conference on Languages and the Media*. Berlin.

Niebaum, Hermann. 1983. *Dialektologie*. Tuebingen: Niemeyer.

Nieuweboer, Rogier. 1998. *The Altai Dialect of Plaudittsch (West-Siberian Mennonite Low German)*. University of Groningen.

Petyt, K.M. 1980. *The Study of Dialect : An Introduction to Dialectology*. London: André Deutsch.

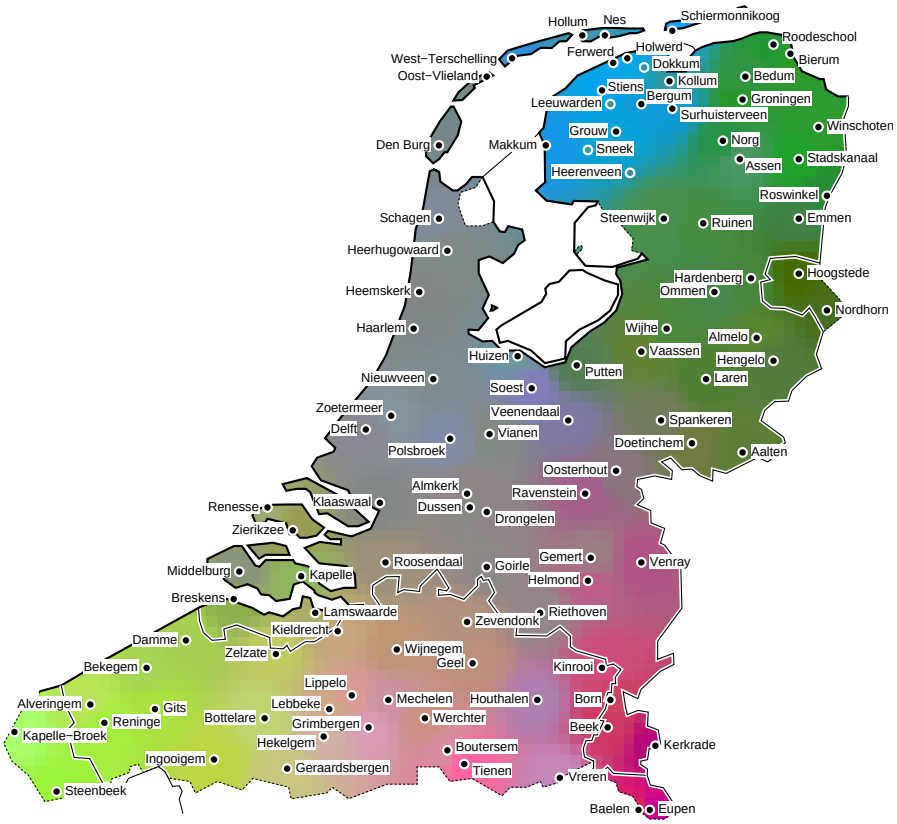
Veldhuijzen van Zanten, Gert, Gosse Bouma, Khalil Sima'an, Gertjan van Noord, and Remko Bonnema. 1999. Evaluation of the NLP Components of the OVIS2 Spoken Dialogue System. OVIS Technical Report 84. Amsterdam/Eindhoven/Groningen: NWO Language-Speech Technology Programme. avail. at `http://odur.let.rug.nl:4321/publijst.html`.

Veronis, Jean. to appear, 2000. *Parallel Text Processing*. Dordrecht and Boston: Kluwer Academic.

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*, ed. Marcel P.R. van den Broecke and A. Cohen, 654–659. Dordrecht: Foris.

Waterman, Michael S. 1995. *Introduction to Computational Biology: Maps, Sequences and Genomes*. London: Chapman and Hall.

Waterman, Michael S. 1989. Sequence Alignments. In *Mathematical Methods for DNA Sequences*, ed. Michael S. Waterman. 53–92. Boca Raton: CRC.



Edit Distance and Dialect Proximity : A Long-standing Problem

A long standing problem in dialectology is the determination of the linguistic basis for dialect areas. Individual words and pronunciations vary irregularly with respect to area boundaries as determined by experts.

The differences in pronunciation of the 100 Dutch words were measured using Levenshtein distance, and the resulting averages were subjected to multi-dimensional scaling to determine the most significant dimensions of difference. These dimensions are colored red, green and blue above, and the color of the intermediate points is determined using inverse distance weighting. The resulting map gives form to the dialectologist's intuition that dialects exist "on a continuum," within which, however, significant differences emerges. The traditional distinctions among the Frisian dialects (blue), Saxon (dark green), Limburg (red), and Flemish (yellow-green) emerge clearly. See the introductory article, "Edit Distance and Dialect Proximity."