University of Pennsylvania Working Papers in Linguistics

Volume 12
Issue 2 *Papers from NWAV 34*

Article 12

1-1-2006

Clustering dialects automatically: A mutual information approach

Naomi Nagy

Xiaoli Zhang

George Nagy

Edgar W. Schneider

Clustering dialects automatically: A mutual information approach

Clustering Dialects Automatically: A Mutual Information Approach

Naomi Nagy, Xiaoli Zhang, George Nagy,
and Edgar W. Schneider

1 Introduction

Dialects can be categorized in many ways. Using external features, dialects may be grouped by geographic location (e.g. Irish English), ethnic identity (e.g. AAVE), or social networks (e.g. Liberian Settler English) of their speakers. Or, using internal features, dialects may be grouped by shared features of pronunciation, vocabulary, or grammar. We explore quantitative approaches to see how similarly dialects cluster by these different methods.

We describe a method of clustering dialects according to patterns of shared phonological features. While previous linguistic research has generally considered such phonological features as independent of each other, we examine their statistical co-variation. That is, we look at the degree to which variation in one feature predicts variation in each other feature, or *Mutual Information* (MI). As an example, we look at the degree to which we can predict whether a dialect will exhibit the *cot/caught* merger based on knowledge of whether they vocalize /r/ in the word *barn*. Within phonological theory, these variables are independent of each other, but they do exhibit statistical dependence.

To test our method, we explore a data set consisting of 168 binary features describing the pronunciation of vowels and consonants of English speakers from 35 countries and regions. This is a subset of the data collected for the *Handbook of Varieties of English* (Schneider et al. 2005). These dialects are grouped according to patterns of shared features. The results of this method of categorizing dialect varieties by binary pronunciation features are compared to traditional groupings based on external features. In many ways, the clusters produced by this method are similar. We also compare differences in clustering outcomes determined by phonological vs. morphosyntactic features, as well as differences that depend on the method of clustering.[1]

[1]We gratefully acknowledge the many contributions to this paper by Benedikt Szmrecsanyi, from first suggesting that we compare analyses to patiently providing many versions of said analyses, without which this paper wouldn't have been possible. The first author also thanks Steve Kirby for stimulating discussions about the

2 Previous work in dialect clustering

There is a (fuzzily) nested set of ways of speaking which, at one extreme of granularity, includes language families such as Germanic or Romance and, at the other end, consists of idiolects. In between, we find languages (e.g. English, German) and dialects (e.g. Midwestern American English), with no clear linguistic distinction between these two. Clustering techniques allow one to look at different sized groupings of linguistic varieties within (or across) languages.

There have been several previous attempts at categorization of dialects. Carver (1987) describes varieties of American English in terms of lexicon and Labov, Ash, and Boberg (2005) do so in terms of phonology. Hughes and Trudgill (1987) and Trudgill (1999) describe the dialects of British English. The aforementioned do not attempt quantified categorization. Recently, there have been sophisticated quantitative analyses of English dialect data (Nerbonne and Kleiweg 2003, Szmrecsanyi and Kortmann 2005), and other languages, e.g. Dutch, Norwegian, Chinese (Cheng 1997, Gooskens and Heeringa 2004, Heeringa 2004, Heeringa and Braun 2003, Heggarty to appear), including some cluster analyses. None of these, however, consider the interrelationships of the phoneme variants across dialects. In this way, our approach is novel.

3 Methods: Data Collection and Organization

The database we are working with is a byproduct of a recent major publication: *A Handbook of Varieties of English* (*HVE*) (Schneider et al. 2005) which describes the pronunciation variants of English in a great many varieties (national, regional, and ethnic) from around the globe (see lists in Appendix and Nagy 2005). The database consists of a spreadsheet with possible pronunciation variants as rows, language varieties as columns, and information on whether or not the respective variant occurs in a given variety as cell entries. A sample of the database is shown in Table 1, which gives the feature frequencies in a cluster of 13 dialect varieties for two phonemes.[2] The first phoneme has 3 allophones or variants, the second has 2. Other possible variants are never realized by speakers in this dialect cluster. In each of the

statistics. We gratefully acknowledge the contribution of the morphosyntactic analysis by Benedikt Szmrecsanyi, building on Szmrecsanyi and Kortmann (2005).

[2] In order to simplify the example calculations in Section 4, we have taken liberties with the data in this table. These are NOT the values reported in the *Handbook*.

5,880 (168x35) feature-by-variety cells, one of three codes originally appeared indicating that in the respective form of English, the respective feature is used (A) regularly, (B) in specific circumstances, or (C) not at all. For the present analysis, binary features are used. "1" indicates that the variant is used regularly (originally A) while "0" indicates that it is used either sometimes (B) or never (C).

VARIETY	KIT			DRESS	
	glide	central	raised	central	raised
Orkney & Shetland		1		1	
North of England			1	1	
East Anglia			1		1
Philadelphia			1		1
Newfoundland	1			1	
Cajun English			1		1
Jamaican Creole		1		1	
Tobago Basilect			1	1	
Australian Creole			1		1
Tok Pisin		1		1	
Fiji English	1			1	
Nigerian Pidgin			1	1	
Indian S. African E.		1		1	
Total	2	4	7	9	4

Table 1: (Imaginary) feature frequencies for 2 words in 13 dialects

To construct this database, Schneider devised a scheme of distinct descriptive categories. He set up a list of 179 features (vowel, consonant, and prosodic features) intended to represent the entire range of possible variants, each of which may or may not be used in each of the varieties under consideration. The list of vowel features builds upon the lexical sets devised by Wells (1982), a system of distinct vowel types identified by certain key words (e.g. TRAP for the vowel in *cat* and *bad*; FACE for the vowel in *rain* or *gate*). 28 different lexical sets are considered, and for each of these, 2-7 different variants are suggested by specifying articulatory features and IPA characters. Table 1 shows five features. They are 3 (of 4) possible variants of KIT: (1) offgliding [iə/ɪə], (2) centralized [ə], and (3) raised and fronted [i]; and 2 (of 6) possible variants of DRESS: half-open [ɛ] and raised [i]. The 121 vowel features can be grouped together in 28 coherent sets of alternative realizations. Within each set, at least one variant should be considered the norm in each variety under consideration. However, the variants are not

mutually exclusive: in many communities more than one variant occurs frequently. Many vowel distribution features relate to mergers, i.e. the fact that certain vowels sound alike (e.g. feature 131 applies if there is homophony between the vowels of LOT and STRUT). Consonant features include a tendency to delete word-initial /h/, and the rhotic realization of postvocalic /r/. The last group includes prosodic features, like the deletion of word-initial unstressed syllables (e.g. *'bout*, *'cept*) or the "high-rising terminal" intonation contour.

The authors of the *HVE* chapters were asked to fill out the list of variants for their respective regions, i.e. to specify for each feature whether or not it occurs. Editors completed feature lists as necessary. Altogether, the columns of the database represent 59 distinct varieties of English, divided into five major world regions. Here, we focus on analyses of the 35 varieties which are included in both the Morphosyntax and the Phonology sections of *HVE* so that comparisons are possible. The geographic distribution and the types of phonological features examined are listed in Table 2. Analyses of similar types, but for a data set containing only phonological data from 59 varieties, were presented in Nagy, Zhang et al. (2005).

Feature type	# features	# variants
vowel	28	121
vowel distributions	4	4
consonants	32	38
prosody	5	5
(omitted	*11)*	
TOTAL	69	168

Geographic distribution	
Africa	9
Americas/Caribbean	9
Asia	3
British Isles	6
Pacific/Aust/NZ	8
TOTAL	35

Table 2: Summary of phonological data

4 Methods: Clustering and Mutual Information

The spreadsheet is analyzed as a binary observation array W, where each element w_{ij} corresponds to a variant of a phonological feature for variety V_i. There are 69 phonological features F_i (six shown in Table 1), with 2-7 variants or possible values per feature. Thus, each binary feature vector w_i has 168 elements. Varieties with 1's in the same column of the array pronounce a given word in the same way, therefore an appropriate measure of the similarity of two varieties V_i and V_j is the Euclidean distance between them. The dissimilarity r_{ij} between two varieties is thus

(1) $r_{ij} = (w_i - w_j)(w_i - w_j)'$

Our starting point for grouping varieties to form dialect clusters is a 35×35 element dissimilarity matrix M. We performed clustering with the Complete Link, Single Link, and Average Link Algorithms (Schütze 2005), which can be found in many statistical data analysis packages (Jain and Dubes 1988). The resulting clusters are mutually exclusive and completely exhaustive: at any given threshold, every variety belongs to exactly one cluster.

A dialect cluster is the *context* that determines the variant (allophone) of each phoneme used by speakers of that dialect. We quantify context by *Mutual Information* (MI), an information theoretic measure calculated from the joint and marginal probability distributions of the allophones of every pair of phonemes. MI is greatest when there is large and consistent variation among the phonological values of the varieties of the cluster. The highest value of MI among two phonemes arises when their variants are all equally probable (and therefore most unpredictable in an information-theoretic sense) among the varieties, and statistically perfectly dependent. Perfect dependence means that knowing how a speaker pronounces one phoneme suffices to predict what variant of the other phoneme will be used by that speaker. For context to be useful, there must be both diversity and dependence across dialects. If all the varieties within a dialect cluster are phonologically similar, then there is no useful context: how speakers pronounce one phoneme reveals nothing about how they pronounce another. Nor is there any useful context if the different speakers' phonological characteristics are statistically independent. This notion can be extended beyond pairs to any number of features, and to any number of varieties.

The result of our analysis is a hierarchy of English dialect clusters with a measure of the MI for the 35 varieties as one cluster, contrasted with the MI found within each cluster when the varieties are clustered into six groups.

The amount of context at any given level of the cluster hierarchy is given by the average MI between pairs of features, for the varieties in that cluster. This measure is based on the marginal and joint probabilities of the features within a cluster. It is equal to the relative entropy between the two distributions: it indicates how much each distribution reveals about the other. MI can represent non-linear statistical dependence, unlike the correlation coefficient. Its formula is:

(2) $I_{x,y} = H(x) - H(x \mid y) = H(y) - H(y \mid x) = \sum_{x,y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$

where p(x,y) is the joint probability distribution of features x and y, and p(x), p(y) are their marginal distributions. H(x) and H(y) are marginal entropies, and H(x|y) is the conditional entropy.

To illustrate, we use the feature frequencies from Table 1. The Mutual Information $I_k(j,l)$ for a pair of phonological features F_j and F_l over all varieties in dialect cluster k at level K is given in (3), where $F_{j,m}$ is the m^{th} variant of the j^{th} feature of variety V_i in dialect cluster C_k.

$$(3)\quad I_k(j,l) = \sum_m \sum_n p(F_{j,m}F_{l,n} \mid V_i \in C_k) \log_2 \frac{p(F_{j,m}F_{l,n} \mid V_i \in C_k)}{p(F_{j,m} \mid V_i \in C_k)\, p(F_{l,n} \mid V_i \in C_k)}$$

Table 3 shows the joint frequency ($p(F_{j,m}\ F_{l,n}|V_i \in C_k)$) and marginal frequencies ($p(F_{j,m}|V_i \in C_k)$ and $p(F_{l,n}|V_i \in C_k)$) of the features in Table 1. The six individual components of MI are shown below: they sum to 0.35.[3]

			KIT		
			glide	central	raised
			0.15 (2/13)	0.31 (4/13)	0.54 (7/13)
DRESS	central	0.69 (9/13)	0.15	0.31	0.23
	raised	0.31 (4/13)	0.00	0.00	0.31

$I(x_i,y_j)=$			
	0.08	0.16	-0.16
	0.00	0.00	0.27

Table 3: Calculating the joint and marginal frequencies for two words in 13 dialects

5 Results: Clustering

Table 4 shows the results of clustering (using real data). This method, using only internal features, constructs clusters that are very similar to those constructed by more traditional dialectology approaches, using both internal and external features. The table allows for exploration of *co-association,* the amount of similarity between the clusters constructed by different methods (Topchy, Jain, and Punch 2004). We can compare two different clustering techniques, Complete Link in column a vs. Average Link in columns b and

[3] $I_{DRESS,KIT}=0.35 < H(x) = 0.89 < \log_2 2 = 1.00;\ H(y) = 1.41 < \log_2 3 = 1.58$

c, and two different sets of observations, Phonology in columns a and b vs. Morphosyntax in column c. The dendrogram in the Appendix illustrates a full cluster analysis and spells out abbreviations used in the text. Other results are available in Nagy, Zhang et al. (2005) and Szmrecsanyi and Kortmann (2005).

The six clusters shown in Table 4 are linguistically highly meaningful, even thrilling; the mathematical procedure yields neatly delimitated, coherent sociohistorical groups of language varieties. What is most interesting is that in a number of instances the results emphasize historical relationships rather than geographical proximity. The clearest case in point is cluster 2 (columns a,b), which unites the southern hemisphere varieties (Australia, New Zealand, South Africa) with East Anglia, a result which lends strong support to the claim that the latter is the primary source of the former (Lass 1987; Trudgill 2004). Cluster 1 brings out the Englishes of South and Southeast Asia (or, for Indian South African English, their descendants) as a closely related group. Cluster 6 (column a) / 2 (columns b,c) models the transmission of English to North America, uniting American English with Irish English and the dialect of Newfoundland. Interestingly enough, two ethnic contact dialects of North America (Chicano English, AAVE) are also shown to be closely related in this group. Cluster 5 combines a Celtic connection in the North and West of the UK (Orkney and Shetlands, Wales) with Scottish English (in column a, and in a different cluster but close by in columns b,c).

Some of the clusters show the effect of language contact quite coherently. Cluster 1 (columns a,b,c) unites almost all varieties that have undergone heavy contact, including pidgins and creoles. It highlights contact-induced similarities from regions as diverse as the Pacific (Hawaii, Vanuatu, Papua New Guinea, Fiji), West Africa, East and South Africa, Australia and the Caribbean. Varieties which historically were produced by even stronger contact and restructuring are singled out in Cluster 3, however: Jamaican, Australian and Surinam creoles.

In future work we will explore a measure of *co-association* to support our sense that there are more differences between the clusterings created from different observations (b,c) than from different clustering techniques (a,b).

	Complete Link	Average Link	
	Phonology		Morphosyntax
	a	b	c
1	Bislama, TP, NigP, GhE, GhP, BlSAfE, InSAfE, PakE, SgE, MalE, CamPE/K CamE, T&TC, HawC, FijiE, EAfE	Bislama, TP, NigP, GhE, GhP, BlSAfE, InSAfe, PakE, SgE, MalE, CamPE/K CamE, T&TC, HawC, FijiE, EAfE AbE	Bislama, TP, NigP, GhE, GhP, BlSAfE, InSAfE, PakE, SgE, MalE, CamPE/K EAfE SurC, WhSAfE
2	WhSAfE, NE, EA, NZE, AusE	WhSAfE, NE, EA, NZE, AusE StAmE, NfldE, AAVE, ChcE, BahE IrE	ColAmE,[4] NfldE, AAVE, ChcE, BahE CamE
3	JamC, AusC SurC AbE	JamC, AusC SurC	JamC, AusC AbE HawC, T&TC, Gullah
4	Gullah BahE	Gullah	OrkS
5	WelE OrkS ScE	WelE OrkS	WelE EA
6	IrE StAmE, NfldE, AAVE, ChcE	ScE	ScE IrE NE, NZE, AusE, FijiE

Table 4: Dialect clusters for two different clustering techniques and two different types of data (K=6)

[4]Different research agendas in the two parts of the *Handbook* necessitate comparing Standard American English phonology data to Colloquial American English morphosyntax data.

6 Results: Mutual Information

While the clustering results illustrate the degree of consistency among dialects, MI shows, whenever there is variation across dialects, how statistically dependent the dialects are on each other. MI can be seen as an additional type of measure, besides similarity, that is of value in distinguishing dialects.

Table 5 lists the amount of MI between each pair of phonemes in a subset of 8 vowels, with all 35 dialects together in one cluster.[5] The 4 highest values are outlined. The dependencies between these vowels are not, to our knowledge, discussed in the dialectology literature. More generally, there is a degree of MI across *every* pair—any word recognition/production application would be improved by including MI in its calculations.

F2 / F1	lax vowels				tense vowels			
	KIT	DRESS	FOOT	THOUGHT	FLEECE	FACE	GOAT	GOOSE
KIT	1.00	0.24	0.54	0.29	0.48	0.43	0.37	0.41
DRESS		1.00	0.16	0.22	0.16	0.16	0.21	0.14
FOOT			1.00	0.29	0.42	0.54	0.46	0.25
THOUGHT				1.00	0.08	0.28	0.23	0.22
FLEECE					1.00	0.50	0.44	0.32
FACE						1.00	0.85	0.36
GOAT							1.00	0.39
GOOSE								1.00

Table 5: Normalized MI for 4 tense and 4 lax vowels, 35 dialects (K=1)

Table 6 shows the value of combining clustering and MI results.[6] This table considers the same 8 words as Table 5, but was calculated for the six clusters shown in Table 4b. Only the 3 largest of the 6 clusters are shown. Shading indicates values for MI that are greater within their cluster than when considering the MI calculated for the 35 dialects as a whole (from Table 5). Over half of the comparisons (the 43 shaded cells, out of 84 total) yield higher MI values. Thus, applications such as voice recognition systems

[5] The values are normalized so that autocorrelations (shaded) = 1. They differ slightly from the example in Table 3, where, for clarity, normalization was not included.

[6] The 0 values indicate a complete lack of variation among the dialects in that cluster for that vowel pair: if there is complete predictability for one of the words, then knowing about the other cannot improve predictions of the first. Auto-comparisons are excluded from this table—their value is always 1.

Phoneme Pairs	T&TC, AbE, BisIm, TP, HawC, FijE, GhE, GhP, CamE, NigP, CamPE/K, EAfE, BlSAfE, MalE InSAfE, PakE, SgE,	IrE, NE, EA, StAmE, NfldE, AAVE, ChcE, BahE, NZE, AusE, WhSAfE	JamC, SurC, AusC
KIT, DRESS	0.27	0.47	0.96
KIT, FOOT	0.43	0.04	0.96
KIT, THOUGHT	0.00	0.12	0.96
KIT, FLEECE	0.42	0.49	0.96
KIT, FACE	0.49	0.25	0.96
KIT, GOAT	0.49	0.43	0.96
KIT, GOOSE	0.16	0.28	0.96
DRESS, FOOT	0.14	0.07	0.26
DRESS, THOUGHT	0.00	0.20	0.96
DRESS, FLEECE	0.21	0.24	0.26
DRESS, FACE	0.24	0.09	0.26
DRESS, GOAT	0.11	0.42	0.26
DRESS, GOOSE	0.08	0.48	0.26
FOOT, THOUGHT	0.00	0.17	0.26
FOOT, FLEECE	0.27	0.03	0.26
FOOT, FACE	0.31	0.05	0.96
FOOT, GOAT	0.20	0.09	0.96
FOOT, GOOSE	0.07	0.07	0.26
THOUGHT, FLEECE	0.00	0.12	0.26
THOUGHT, FACE	0.00	0.33	0.26
THOUGHT, GOAT	0.00	0.59	0.26
THOUGHT, GOOSE	0.00	0.39	0.26
FLEECE, FACE	0.54	0.56	0.26
FLEECE, GOAT	0.24	0.54	0.26
FLEECE, GOOSE	0.16	0.28	0.96
FACE, GOAT	0.59	0.64	0.96
FACE, GOOSE	0.09	0.39	0.26
GOAT, GOOSE	0.05	0.74	0.26

Table 6. Normalized MI for 4 tense and 4 lax vowels, for 3 largest dialect clusters (K=6)

would be improved by individually trained classifiers for each dialect cluster. This finding is in keeping with what has been shown for MI as applied to handprinting recognition (Veeramachaneni and Nagy 2005).

7 Applications and Future Work

We have examined the phonological correlates of English dialects from the orthogonal perspectives of consistency (clustering) and context (MI). Hierarchical clustering organizes dialects with similar pronunciations. MI, on the other hand, reveals a high level of statistical dependence between alternative pronunciations of pairs of vowels within the same dialect cluster. This second aspect is novel. Its value must be assessed by further investigation: dialects are not traditionally characterized by their statistical inter-dependence. Given access to appropriate data, perhaps from Cheng (1997), Gooskens and Heeringa (2004), Heeringa and Braun (2003), we could test the method with other languages.

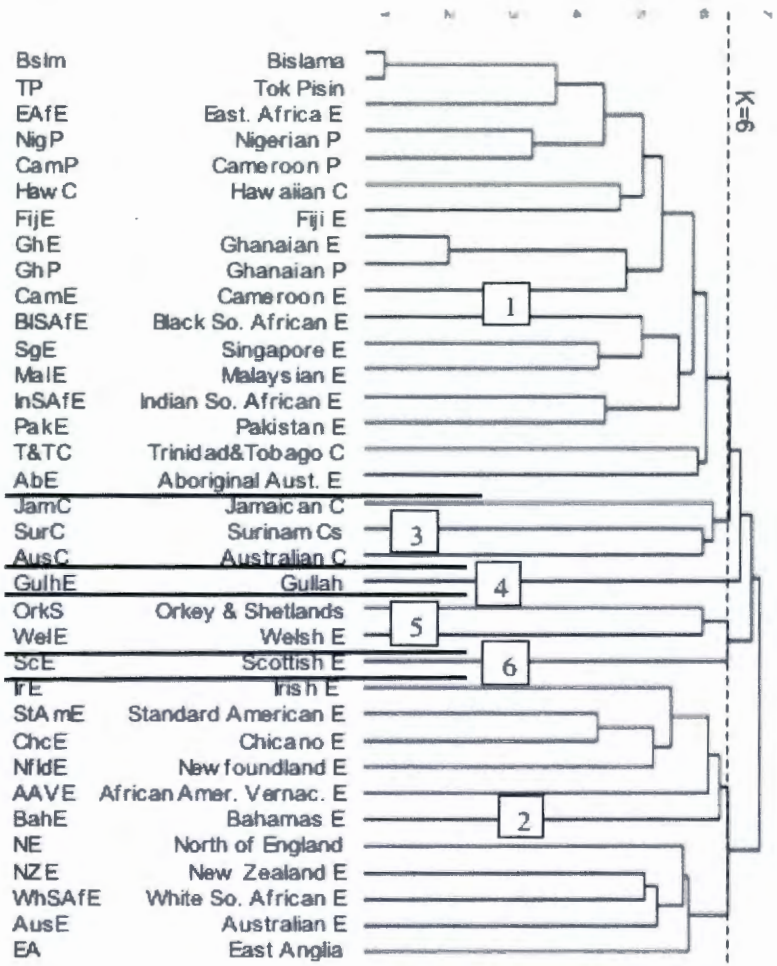
Ideally we would test these methods at all levels of the continuum from idiolect to language. The necessary data would include descriptions of many idiolects for each dialect, just as we have many dialects for the one language considered here. Once such a classification is obtained, we would be able to predict, for a partially unanalyzed dialect, what features it will exhibit based on knowledge of some subset of features that it has been shown to exhibit. This could be applied to speaker identification by permitting a stochastic description of a speaker's full dialect based on a sample which contains only a subset of the phonemes.

Phonological context may also find practical application in automated speech recognition (ASR). This technology has made good progress since the first attempts in the 1960s to recognize "yes" vs. "no" for accepting or declining a collect call. ASR has been deployed for telephone trees, directory assistance, and queries for stock-market prices. Other restricted-vocabulary dialogs, for airline reservations and for hands-free operations like stock inventory and non-critical vehicular applications (radio, seat adjustment, cellphone dialing), have also been developed. Large-vocabulary trainable dictation systems have been available for several years. In most of these applications, recognition accuracy could be raised by exploiting both the consistency and the statistical dependencies in the pronunciation of speakers within a given dialect cluster.

One caveat is that this will be useful only if it can be verified from acoustic waveforms that most of the speakers of a variety actually pronounce the words in the ways that have been described, and if that can be reliably detected automatically. Multi-modal Hidden Markov Models, widely used in

speech recognition (Rabiner and Juang 1993), would provide the appropriate framework for continuing this work with automated phonological characterization. Further interdisciplinary studies could render differences between dialects an advantage, rather than a detriment, to ASR.

Appendix: Abbreviations, Dendrogram for Table 4b



This dendrogram was created using the Average Link Method-Euclidean distance, phonological data, for 35 varieties, K=6. It corresponds to the clusters shown in Column b of Table 4.

References

Carver, Craig M. 1987. *American Regional Dialects: A Word Geography*. Ann Arbor: University of Michigan Press.

Cheng, Chin-Chuan. 1997. Measuring Relationship among Dialects: DOC [Dictionary on computer] and Related Resources. *Computational Linguistics and Chinese Language Processing* 2.1:41–72.

Gooskens, Charlotte and Wilbert Heeringa. 2004. Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. *Language Variation and Change* 16.3:189–207.

Heeringa, Wilbert. 2004. Measuring dialect pronunciation differences using Levenshtein distance. Groningen: University of Groningen.

Heeringa, Wilbert and Angelika Braun. 2003. The Use of the Almeida-Braun System in the Measurement of Dutch Dialect Distances. *Computers and the Humanities* 37.3:257–271.

Heggarty, Paul A. *to appear. Measured Language: From First Principles to New Techniques for Putting Numbers on Language Similarity*. Oxford: Blackwell.

Hughes, Anne and Peter Trudgill. 1987. *English Accents and Dialects: An Introduction to Social and Regional Varieties of British English.* London: Edward Arnold.

Jain, Anil K. and R.C. Dubes. 1988. *Algorithms for Clustering Data*: Prentice Hall.

Labov, William, Sharon Ash, and Charles Boberg. 2005. *Atlas of North American English*. Paris: Mouton de Gruyter.

Lass, Roger. 1987. *Where do extraterritorial Englishes come from? Dialect input and recodification in transported Englishes*. Fifth International Conference on English Historical Linguistics, Benjamins.

Nagy, Naomi. 2005. Addenda to Categorization of phonemic dialect features in context. *http://pubpages.unh.edu/~ngn/papers/Context05/CONTEXT05_addenda.*

Nagy, Naomi, Xiaoli Zhang, George Nagy, and Edgar Schneider. 2005. A Quantitative categorization of phonemic dialect features in context. In *CONTEXT 2005, Lecture Notes in Artificial Intelligence 3554*, eds. Anind Dey, Boicho Kokinov, David Leake and Roy Turner, 326–338. Berlin / Heidelberg: Springer-Verlag.

Nerbonne, John and Peter Kleiweg. 2003. Lexical distance in LAMSAS. *Computers and the Humanities* 37.3:339–357.

Rabiner, Lawrence R. and Bing-Hwang Juang. 1993. *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice Hall.

Schneider, Edgar W., Kate Burridge, Bernd Kortmann, Rajend Mesthrie, and Clive Upton, eds. 2005. *A Handbook of Varieties of English: A Multimedia Reference Tool*. Berlin, New York: Mouton de Gruyter.

Schütze, Hinrich. 2005. Single-Link, Complete-Link & Average-Link Clustering. *http://www-csli.stanford.edu/~schuetze/completelink.html*. Accessed 2/26/2005.
Szmrecsanyi, Benedikt and Bernd Kortmann. 2005. *The quest for angloversals and vernacular universals in varieties of English world-wide*. NWAV34, New York.
Topchy, Alexander, Anil K. Jain, and William Punch. 2004. *A Mixture Model for Clustering Ensembles*. Proc. SIAM International Conference on Data Mining, Florida.
Trudgill, Peter. 1999. *The Dialects of England*. London: Blackwell.
Trudgill, Peter. 2004. *New Dialect Formation. The Inevitability of Colonial Englishes*. Edinburgh: Edinburgh University Press.
Veeramachaneni, Sriharsha and George Nagy. 2005. Style context with second order statistics. *IEEE Trans. Pattern Analysis and Machine Intelligence* 27(1):14–22.
Wells, John C., ed. 1982. *Accents of English*. Cambridge: Cambridge University Press.

Naomi Nagy
English Department
University of New Hampshire
Durham, NH 03824 USA
ngn@unh.edu

Xiaoli Zhang, George Nagy
DocLab, ECSE
Rensselaer Polytechnic Institute
Troy, NY 12180 USA
{zhangxl, nagy}@rpi.edu

Edgar Schneider
Department of English Linguistics
Regensburg University
Regensburg, Germany
edgar.schneider@sprachlit.uni-regensburg.de