# A Quantitative Categorization of Phonemic Dialect Features in Context

Naomi Nagy[1], Xiaoli Zhang[2], George Nagy[2], and Edgar W. Schneider[3]

[1] English Department, University of New Hampshire, Durham, NH 03824 USA
ngn@unh.edu

[2] DocLab, ECSE, Rensselaer Polytechnic Institute, Troy, NY 12180 USA
{zhangxl, nagy}@rpi.edu

[3] Department of English Linguistics, Regensburg University, Regensburg, Germany
edgar.schneider@sprachlit.uni-regensburg.de

**Abstract.** We test a method of clustering dialects of English according to patterns of shared phonological features. Previous linguistic research has generally considered phonological features as independent of each other, but context is important: rather than considering each phonological feature individually, we compare the patterns of shared features, or *Mutual Information (MI)*. The dependence of one phonological feature on the others is quantified and exploited. The results of this method of categorizing 59 dialect varieties by 168 binary internal (pronunciation) features are compared to traditional groupings based on external features (e.g., ethnic, geographic). The MI and size of the groups are calculated for taxonomies at various levels of granularity and these groups are compared to other analyses of geographic and ethnic distribution. Applications that could be improved by using MI methods are suggested.

## 1 Introduction

The way a given language is spoken by a particular group at a particular time is referred to as a dialect. Dialects can be grouped into categories in many different ways. Using *external features*, dialects may be grouped by geographic location (*e.g.*, Irish English), ethnic identity (*e.g.*, African-American Vernacular English), or social networks (*e.g.*, Liberian Settler English) of their speakers. Alternatively, 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.

How many dialects are there? For English, answers may range from 1 to ~341,000,000 (the number of mother-tongue speakers of English). Dialects can be as narrow as that of a single speaker (an *idiolect*), or as broad as a major language (*e.g.*, Chinese or Spanish). While no two speakers speak identically, speakers may be grouped into dialects by degree of similarity of vocabulary (lexical categories), grammar (syntactic categories), and/or pronunciation (phonological categories). We exploit a data set showing variations in the pronunciation of a set of vowels and consonants to form dialect clusters by phonological categorization. From this perspective, a *dialect cluster* is the context that determines the variant (*allophone*) of each phoneme used by speakers of that dialect. We analyze pronunciation, rather than lexicon, in part because a more robust classification system should be possible due to the much smaller number (and concomitant higher frequency) of sounds than words in any given language.

Context requires both diversity and dependence. If all the varieties within a dialect cluster are phonologically similar, then there is no phonological context: how speakers pronounce one phoneme reveals nothing about how they pronounce another. Nor is there any context if the different speakers' phonological characteristics are statistically independent. While "context" has a broad range of definitions in logic, linguistics, philosophy, artificial intelligence, and sociology, inter alia [1, 2], we chose a narrow definition in order to operationalize it. 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 strongest possible context among two features 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. This notion of context can be extended beyond pairs to any set of features, and to any number of varieties in a cluster.

The result of our analysis is a hierarchy of English dialect clusters with a measure of the MI at each level. Aside from its intrinsic interest in linguistics for comparison with alternative taxonomies, this approach may decrease error rates in automatic speech recognition (ASR) for dialectically homogeneous groups of speakers. Similar methods, based on *style context*, have met with some success in the recognition of hand-written digits and printed text [3, 4]. Whereas in speech the style context is provided by dialect, in hand-print it may be due to each form in a batch filled out by a single writer, and in printed text it may originate from a commonality of font, printer, scanner or copier.

In Sec. 2, we sketch the foundations of phonology, and the formation and characteristics of dialects and their taxonomies. Secs. 3 & 4 present the rationale and collection protocol for our phonological data. Although the clustering algorithm and probabilistic distance measure that we use are not new to computer and information scientists, we illustrate them with brief examples using phonological features. (Our contribution is the adaptation of *hierarchical clustering* and of a *measure of statistical dependence* to new linguistic data.) Secs. 5 & 6 present the groupings we obtained with their information-theoretic measure of context and a comparison of our dialect clusters with groupings obtained by alternative methods. Sec. 7 outlines applications in the domains of digital speech and automatic speech recognition that could be improved by using these methods.

## 2 Phonology, Language and Dialects

At the phonological level, the units of spoken language are phonemes, the smallest units of sound recognized as distinct by speakers of the language. For example, [t] and [d] are distinct phonemes in English—speakers recognize that they distinguish the words "try" and "dry" (as well as many other word pairs, like "lit" and "lid"). Both are articulated as alveolar stops, but the first is voiceless and the second voiced. These three articulatory features (place and manner of articulation and voicing) are used to uniquely identify the consonant phonemes of a language. Finer phonetic distinctions exist: phonemes may be pronounced differently depending on the context in which they are found. These different pronunciations are referred to as allophones of a particular phoneme, e.g., the difference between the aspirated [$t^h$] in "take" vs. the unaspirated [t] in "stake.". Native speakers of a language are rarely conscious of these sub-phonemic or allophonic differences, but they are important dialect markers.

Vowels are described in terms of tongue position, lip rounding and duration. Tongue height and backness, combined with duration, uniquely distinguish the vowels of English. Again, finer-grained distinctions exist below the level of consciousness of speakers. It is these phonetic distinctions that are described by the 168 binary features in our descriptive system.

A new language variety develops when a group of people maintain contact with each other over an extended time period and are isolated from other speakers. A variety which starts out as a dialect of a language may, given enough time, develop into a new language (no longer mutually intelligible). For example, English got its beginnings as Anglo Saxon, a Germanic dialect, when people from (present-day) Germany (Angles, Saxons, and Jutes) settled in (present-day) England. Due to a period of extended isolation from other German speakers, a non-mutually intelligible dialect eventually developed.

What is of importance here is that 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, includes idiolect. In between, we find languages (*e.g.*, English, German) and dialects (*e.g.*, Midwestern American English), with no clear-cut distinction between these two. In this paper we look at different size groupings of linguistic varieties within the English language.[1]

There have been several previous attempts at categorization of dialects. [5] describes varieties of American English in terms of lexicon and [6] does so in terms of phonology. [7] and [8] describe the dialects of British English. The aforementioned do not attempt quantified categorization. Recently, there have been sophisticated quantitative analyses of English dialect data [9], and other languages (Dutch, Norwegian, Chinese) [10-14], including some cluster analyses. None of these, however, consider the interrelationship of the phoneme variants across dialects.

[1] "Linguistic variety" is a cover term for idiolects, dialects, and languages.

## 3 Methods: Data Collection and Organization

Our database is a side product of a major publication project: *A Handbook of Varieties of English* [15] which describes the pronunciation variants of English in a great many varieties (national, regional and ethnic dialects) from around the globe (see list in [16]). 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 (see partial example in Table 2).

For publication, the pronunciation variation needed to be systematized in order to produce categorical displays of certain phenomena as realized in specific varieties. To allow tabular and cartographic representations of the essentially infinite pronunciation variability world-wide, E.W. Schneider devised a scheme of distinct descriptive categories. One difficulty in this process was that the range of possible variants was not fully known in advance; another was to decide how finely to sub-categorize in order to remain both informative and descriptively adequate as well as manageable. Schneider set up a listing of 179 features of pronunciation (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 (see Table 1).

The list of *vowel features* builds upon the "lexical sets" devised by [17], 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 realization possibilities (variants) are suggested by specifying articulatory features and International Phonetic Alphabet characters. For example, in features 1-4, possible variants of the vowel of KIT are identified as (1) "canonical" high front [ɪ]; (2) raised and fronted variant phonetically identified by the symbol [i], (3) centralized [ə], and (4) with an offglide, *e.g.* [ɪə/iə]. Thus, the 121 vowel features can be grouped together in 28 coherent sets of alternative realizations. At least one of these variants should apply to each of the language varieties under consideration. However, the variants need not be mutually exclusive: in many communities the degree of variability is high and more than one variant may occur. The *vowel distribution features* relate to so-called mergers, *i.e.*, the fact that certain vowel types sound alike (so, for instance, feature 131 applies if there is homophony between the vowels of LOT and STRUT). The *consonant features* include a tendency to delete word-initial *h*- ('*eart* 'heart'), or the rhotic realization of postvocalic /r/ ([bɑɹn] vs. [ba:n] ‘barn’). The last group includes *prosodic features*, like the deletion of word-initial unstressed syllables (*e.g.* '*bout*, '*cept*) or the "high-rising terminal" contour, a tendency to raise one's pitch at the end of declarative statements.

The authors of the Handbook chapters were asked to fill out the list of pronunciation variants for their respective regions, *i.e.,* to specify for each feature whether or not it occurs. To achieve a roughly even coverage of varieties, the regional editors filled in the feature lists as necessary. Altogether, the columns of the database used here represent 59 distinct varieties of English, divided into four major world regions. (See Table 1.)

**Table 1.** Summary of phonological data

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

| *Geographic distribution* | |
|---|---|
| British Isles | 9 |
| Pacific & Australia | 10 |
| Africa & Asia | 19 |
| Americas & Caribbean | 21 |
| TOTAL | 59 |

In each of the 10,561 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 statistical 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).

## 4 Methods: Clustering and Mutual Information

The completed data sheets described above were transformed into 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 $F_i$ (See Table 1), with 2-7 variants or possible values per feature. Thus, each binary feature vector $\mathbf{w}_{\mathrm{i}}$ 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 number of 1's in the logical AND of their feature vectors, normalized by the product of their lengths. (See Table 3 below.) Then, the *dissimilarity* $□_{ij}$ between two varieties is

$$□_{ij} = 1 - |\mathbf{w}_{\mathrm{i}} □ \mathbf{w}_{\mathrm{j}}|/|\mathbf{w}_{\mathrm{i}}|\ |\mathbf{w}_{\mathrm{j}}| = 1 - \cos\,(\mathbf{w}_{\mathrm{i}}\ \mathbf{w}_{\mathrm{j}})\ . \tag{1}$$

Our starting point for grouping varieties to form *dialect clusters* is a 59□59 element *dissimilarity matrix M*. We note that there is no general way of determining, from a similarity or dissimilarity matrix or from an array of feature vectors, how many clusters there are in a data set. Clustering, or "unsupervised learning," requires some external information, such as the maximum acceptable distance between patterns in the same cluster, or the minimum distance between patterns from different clusters, or the minimum or maximum number of patterns in a cluster. Clustering may be *hierarchical* or *flat*, *agglomerative* or *divisive*, and *crisp* (mutually exclusive clusters) or *fuzzy* (with a continuous cluster membership function). Dozens of clustering algorithms have been developed and applied [18-21]. Objects characterized by a similarity matrix can be transformed into a vector space by multi-dimensional scaling; conversely, the similarity of feature vectors can be obtained from their pairwise distance [22]. Current research focuses on *clustering ensembles*, *i.e.*, on combining the results of diverse clustering algorithms [23, 24], and on related algorithms for probabilistic Expectation Maximization [25, 26].

We performed clustering with the Complete Link Algorithm (hierarchical, agglomerative, and crisp), which can be found in many statistical data analysis packages [20]. At any given threshold, the Complete Link Algorithm forms clusters such that the maximum dissimilarity between any two varieties in the cluster is less than □. Clusters are merged when the maximum dissimilarity between a variety in one cluster and a variety in the other cluster is less than □. The resulting clusters are mutually exclusive, and completely exhaustive: at any given threshold, every variety belongs to exactly one cluster.

Initially, each variety is a distinct dialect cluster (an idiolect). The threshold is increased from 0 to 1 to decrease the number of clusters. At 1 (or at any value of the threshold greater than the dissimilarity of the least similar pair of varieties), all the varieties are merged into a single dialect (the English language). A simple example is given in [16].

The amount of context at level □ of the hierarchy is given by the average MI between pairs of features. 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:

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

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, Table 2 shows the feature frequencies in a dialect cluster of 13 varieties for two phonemes. The first phoneme has three allophones, the second has two.

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

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

where $F_{j,m}$ is the m$^{th}$ variant of the j$^{th}$ feature of variety $V_i$ in dialect cluster $C_k$.

Table 3 shows the joint frequency ($p(F_{j,m}\ F_{l,n}|V_i$ □ $C_k)$) and marginal frequencies ($p(F_{j,m}|V_i$ □ $C_k)$ and $p(F_{l,n}|V_i$ □ $C_k)$) of the two features. The six individual components of MI are shown below: they sum to *0.35*.[2]

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

**Table 2.** Feature frequencies for two words in 13 dialects of English

| VARIETY | KIT | | | DRESS | |
|---|---|---|---|---|---|
| | *raised* | *central* | *back* | *raised* | *central* |
| 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 Eng. | 1 | | | 1 | |
| Tok Pisin | | 1 | | | 1 |
| Fiji English | | | 1 | | 1 |
| Nigerian Pidgin | 1 | | | | 1 |
| Indian S. African Eng. | | 1 | | | 1 |
| **Total** | **7** | **4** | **2** | **4** | **9** |

**Table 3.** Calculations of joint and marginal frequencies for two words in 13 dialects of English

| | | | KIT | | |
|---|---|---|---|---|---|
| | | | *back* | *central* | *raised* |
| | | | 0.15 | 0.31 | 0.54 |
| **DRESS** | *central* | 0.69 | 0.15 | 0.31 | 0.23 |
| | *raised* | 0.31 | 0.00 | 0.00 | 0.31 |

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

Since many previous dialectology studies focused on vowel features [27], we also calculated MI separately for the 28 vowel features and the 32 consonant features.

## 5 Results: Clustering

The dendrogram and tables in this section show the results of clustering at various levels. We show that this method, using only internal features, constructs clusters that are very similar to those that have been constructed by more traditional dialectology approaches, using both internal and external features. Fig. 1 shows the clusters achieved with varying thresholds, using all features. A horizontal line marks K=10, the threshold used in discussion below. Table 4 lists the dialects in each cluster, and provides the thresholds (□) at which the dialects fall into 10 clusters. (See [16] for dialect names.) Similar clusters were achieved using the subsets of only vowels and then only consonants, and at K=20 and K=30.

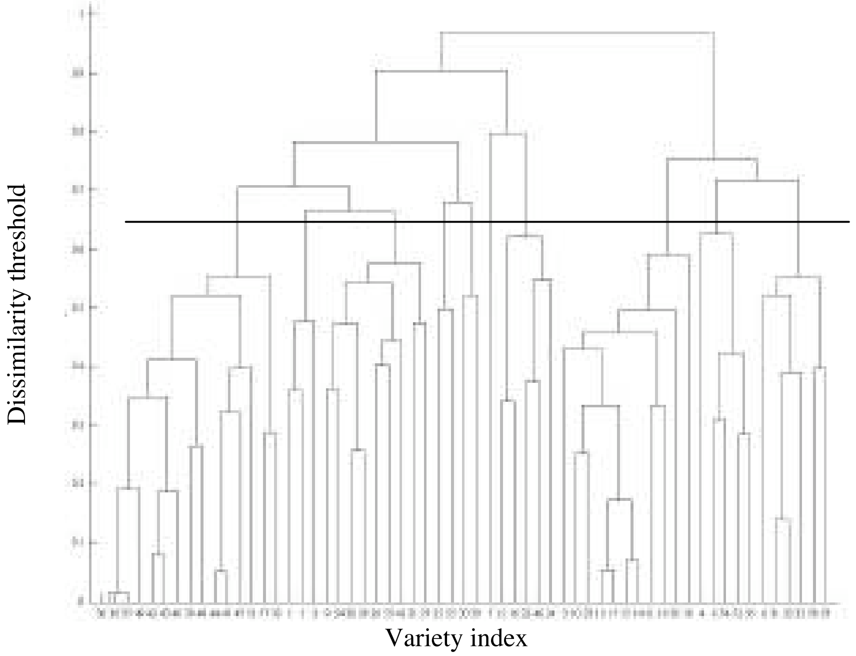

**Fig. 1.** Dendrogram of Complete Link Analysis for all features (horizontal line at K = 10)

The resulting clusters are remarkably meaningful and homogeneous in a linguistic perspective. Based upon the analysis of all features, at K=10, clusters 1 through 5 are extremely tight-knit, and the following two clusters, while less obvious, also allow meaningful interpretation. Cluster 1, the biggest, comprises the Pacific contact varieties on the one hand (Bislama, Tok Pisin, Solomons Pijin, Hawaiian Creole, Fijian English) and a strong cohort of African pidgins and contact Englishes (of Nigeria, Ghana, Cameroon, East Africa, and black South Africa) on the other. Singaporean and Malaysian English, also strongly contact-shaped, also occur in this group. Cluster 2 unites the oldest colonial offspring of British English: Irish English and most varieties of American English (New England, New York, Philadelphia, Inland Northern, Western), including Canadian English and two American contact dialects, African American Vernacular English (AAVE) and Chicano English. Cluster 3 combines the Caribbean Creoles of Jamaica, Barbados, Trinidad and Tobago, as well as their closest kin in North America, Gullah; and in Britain, British Creole. In addition, a few non-contiguous contact dialects can be found in this cluster: Australian Aboriginal English, northern Nigerian English, and Indian English.

Cluster 4 groups the so-called "southern hemisphere" dialects, namely Australian and New Zealand English, Maori English, white So. African English, and Cape Flats English.

**Table 4.** Dialect clusters for 3 different sets of phonological features (K = 10)

| K=10 | **All features** | **Vowel features** | **Consonant features** |
|---|---|---|---|
| □ | 0.63 | 0.6 | 0.8 |
| 1 | {Bislm, TP, Pijin, HawC, FijE, NigES, NigP, GhE, GhP, CamE, CamPE/K, EAfE, BlSAfE, SgE, MalE} | {Bislm, TP, Pijin, HawC, FijE, NigES, NigP, GhE, GhP, CamE, CamPE/K, EAfE, BlSAfE, SgE, MalE} | {BrC, GulhE, CajE, JamC, T&TC, TobgB, SurC, AbE, AusC, Bislm, TP, Pijin, HawC, FijE, SgE} |
| 2 | {IrE, StAmE, NEngE, InlNE, NYCE, PhilE, WMwE, CanE, NfldE, AAVE, ChcE} | { IrE, StAmE, NEngE, InlNE, NYCE, PhilE, SE, UrbS, WMwE, NfldE, AAVE, CajE, ChcE, BahE, LibSE} | { IrE, StAmE, NEngE, InlNE, NYCE, PhilE, SE, UrbS, WMwE, CanE, ChcE, NZE, MaoE, AusE} |
| 3 | {BrC, GulhE, JamE, BarbE, T&TC, TobgB, AbE, NigEN, IndE} | {NE, BrC, GulhE, JamE, BarbE, AbE, NigEN, IndE} | {NfldE, NigES, NigP, LibSE, CamE, CamPE/K, EAfE, BlSAfE, IndSAfE, StHE, MalE} |
| 4 | {EA, NZE, MaoE, AusE, WhSAfE, CFE} | {Chanl, WhSAfE, IndSAfE, CFE, StHE, PakE} | {Chanl, NigEN, GhE, GhP, WhSAfE, CFE} |
| 5 | {SE, UrbS, CajE, BahE, LibSE} | {OrkS, ScE, WelE, T&TC, TobgB} | {OrkS, WelE, IndE, PakE} |
| 6 | {NE, Chanl, IndSAfE, StHE, PakE} | {EA, NZE, MaoE, AusE} | {ScE, JamE} |
| 7 | {OrkS, ScE, WelE} | {JamC, AusC} | {NE, EA} |
| 8 | {JamC, AusC} | {SurC, PhlE} | {RP, PhlE} |
| 9 | {SurC, PhlE} | {CanE} | {AAVE, BahE} |
| 10 | {RP} | {RP} | {BarbE} |

Interestingly, it combines these with only one other dialect which has frequently been suspected to be a major donor of these colonial varieties, the dialect of East Anglia. Cluster 5 unites dialects with a historical or physical connection to the American South: urban and rural Southern, Cajun, Bahamian, and Liberian Settler English (the variety spoken by descendants of repatriated American slaves). Cluster 7, with Wales, Scotland, and Orkney and Shetland, unites some non-central dialects of Britain. Interestingly enough, RP (standard non-regional British English) stands on its own.

The clustering based on vowel features only is very similar in many respects — not surprisingly, given that this subset covers the majority of the features. In comparison with the "all features" categorization, Cluster 1 is identical. Cluster 2 brings out the unity of American English even more strongly, combining almost all American dialects including Southern. It is noteworthy that Irish English groups with the American dialects in all analyses. The only exception (also meaningful, as this is the only English-based creole on the North American mainland) is Gullah, which still groups with the Caribbean Creoles and other contact varieties, a group which now includes the North of England. One unexpected outcome is that Trinidad and Tobago join the Scottish-Welsh cluster.

Looking at consonants only, the resulting patterns are somewhat different. Cluster 1 unites the Caribbean creoles on the one hand with the Pacific pidgins on the other. Cluster 2 combines long-standing first language colonial varieties: Ireland, most dialects of American English (from Canada to the South, from New England to the West), and the antipodean varieties. Cluster 3 consists mostly of varieties of English as a Second Language with a strong focus on all parts of Africa, but including also Singapore and, not fitting this description, Newfoundland. Cluster 4 has further African varieties, in addition to, surprisingly, the Channel Islands. In cluster 5 we find two pairs which are geographically relatively coherent internally but surprising in their mutual combination: the Orkney and Shetlands and Wales on the one hand, India and Pakistan on the other. Three more clusters with pairs of varieties are interesting linguistically: one with two distinct English dialects (7: North, and East Anglia), one with two historically related dialects (AAVE and Bahamian), and one that puts Scottish and Jamaican English together. The unlikely combinations found when comparing consonants rather than vowels perhaps explains why earlier research has included more discussion of vowels, a context in which external and external features produce similar clusters.

## 6 Results: Mutual Information

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

Table 5 lists the amount of MI between each pair of phonemes in a subset of 8 features (4 tense and 4 lax vowels), with all dialects together in one cluster. Auto-comparisons are shaded. The 3 highest values are outlined—interestingly, all involve the GOAT vowel. These dependencies 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.

**Table 5.** MI for 4 tense and 4 lax vowels, all dialects

| F2 | lax vowels | | | | tense vowels | | | |
|---|---|---|---|---|---|---|---|---|
| F1 | KIT | DRESS | FOOT | THOUGHT | FLEECE | FACE | GOAT | GOOSE |
| KIT | 2 | 0.41 | 0.58 | 0.33 | 0.52 | 0.61 | 0.69 | 0.51 |
| DRESS | | 1.48 | 0.13 | 0.30 | 0.24 | 0.3 | 0.40 | 0.32 |
| FOOT | | | 1.4 | 0.28 | 0.48 | 0.58 | 0.53 | 0.29 |
| THOUGHT | | | | 1.41 | 0.24 | 0.44 | 0.41 | 0.56 |
| FLEECE | | | | | 1.53 | 0.57 | 0.68 | 0.42 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| FACE | | | | | | 2.24 | 1.30 | 0.58 |
| GOAT | | | | | | | 2.33 | 0.57 |
| GOOSE | | | | | | | | 1.56 |

**Table 6.** MI for 4 tense and 4 lax vowels, for 10 dialect clusters

| K = 10,<br>□ = 0.63 | {Bislm TP Pijin HawC FijE NigES NigP GhE GhP CamE CamPE/K EAfE BlSAfE SgE MaleE} | {IrE StAmE NEngE InlNE NYCE PhilE WMwE CanE NfldE AAVE ChcE} | {BrC GulhE JamE BarbE T&TC TobgB AbE NigEN IndE} | {EA NZE MaoE AusE WhSAfE CFE} | {SE UrbS CajE BahE LibSE} | {NE Chanl IndSAfE StHE PakE} |
|---|---|---|---|---|---|---|
| KIT, KIT | 1.16 | 0 | 0.92 | 1.92 | 1.37 | 1.37 |
| KIT, DRESS | 0.57 | 0 | 0.07 | 0.92 | 0.72 | 0.97 |
| KIT, FOOT | 0 | 0 | 0.25 | 0 | 0.17 | 0 |
| KIT, THOUGHT | 0 | 0 | 0.31 | 0 | 0.82 | 0 |
| KIT, FLEECE | 0.47 | 0 | 0.46 | 0.58 | 0.82 | 0 |
| KIT, FACE | 0.09 | 0 | 0.46 | 0.79 | 0.97 | 0.42 |
| KIT, GOAT | 0.04 | 0 | 0.46 | 1.58 | 1.37 | 0.97 |
| KIT, GOOSE | 0.13 | 0 | 0.20 | 0.32 | 1.37 | 0 |
| DRESS, DRESS | 1.55 | 0.44 | 0.50 | 1.25 | 0.72 | 1.52 |
| DRESS, FOOT | 0 | 0.01 | 0.04 | 0 | 0.07 | 0 |
| DRESS, THOUGHT | 0 | 0.11 | 0.5 | 0 | 0.72 | 0 |
| DRESS, FLEECE | 0.24 | 0.44 | 0.04 | 0.71 | 0.72 | 0 |
| DRESS, FACE | 0.11 | 0.01 | 0.04 | 0.46 | 0.32 | 0.17 |
| DRESS, GOAT | 0.05 | 0.03 | 0.04 | 0.92 | 0.72 | 1.12 |
| DRESS, GOOSE | 0.13 | 0.26 | 0.02 | 0.11 | 0.72 | 0 |
| FOOT, FOOT | 0 | 0.44 | 0.99 | 0 | 0.72 | 0 |
| FOOT, THOUGHT | 0 | 0.11 | 0.55 | 0 | 0.17 | 0 |
| FOOT, FLEECE | 0 | 0.01 | 0.53 | 0 | 0.72 | 0 |
| FOOT, FACE | 0 | 0.01 | 0.53 | 0 | 0.32 | 0 |
| FOOT, GOAT | 0 | 0.03 | 0.53 | 0 | 0.32 | 0 |
| FOOT, GOOSE | 0 | 0.06 | 0.50 | 0 | 0.17 | 0 |
| THOUGHT,THOUGHT | 0 | 1.68 | 1.45 | 0 | 1.37 | 0 |
| THOUGHT,FLEECE | 0 | 0.11 | 0.55 | 0 | 0.82 | 0 |
| THOUGHT, FACE | 0 | 0.11 | 0.55 | 0 | 0.57 | 0 |
| THOUGHT, GOAT | 0 | 0.24 | 0.55 | 0 | 0.97 | 0 |
| THOUGHT, GOOSE | 0 | 0.80 | 0.50 | 0 | 0.82 | 0 |
| FLEECE, FLEECE | 1.05 | 0.44 | 0.99 | 1.25 | 1.37 | 0 |
| FLEECE, FACE | 0.38 | 0.01 | 0.99 | 0.46 | 0.57 | 0 |
| FLEECE, GOAT | 0.03 | 0.03 | 0.99 | 0.92 | 0.97 | 0 |
| FLEECE, GOOSE | 0.10 | 0.26 | 0.50 | 0.11 | 0.82 | 0 |
| FACE, FACE | 0.70 | 0.44 | 0.99 | 1.46 | 1.52 | 0.97 |
| FACE, GOAT | 0.01 | 0.03 | 0.99 | 0.79 | 1.52 | 0.97 |
| FACE, GOOSE | 0.05 | 0.06 | 0.50 | 0.65 | 0.97 | 0 |
| GOAT, GOAT | 0.35 | 0.87 | 0.99 | 1.92 | 1.92 | 1.92 |
| GOAT, GOOSE | 0.02 | 0.49 | 0.50 | 0.32 | 1.37 | 0 |
| GOOSE, GOOSE | 0.72 | 1.49 | 0.50 | 0.65 | 1.37 | 0 |

Table 6 shows both clustering and MI results. This table considers the same 8 words as Table 5, but was calculated for K=10. Only the 6 clusters containing 5 or more dialects are shown. Again auto-comparisons are shaded. The 4 outlined cells illustrate the value of combining clustering and MI: these values are all greater within their clusters than for the 59 dialects as a whole (where MI=0.41). Thus, applications such as voice recognition systems would be improved by individually trained classifiers for each dialect cluster. The value of MI is affected by the number of values that occur per feature. This depends both on the selected pair of features, and on the varieties included in a cluster. Note that a really tight cluster would necessarily have low MI values—whenever the dialects share the same features, the variation in features cannot be used to predict patterns.

Table 6 shows that MI provides information useful in predicting pronunciation patterns—there are *no* cases of completely independent variation. Again, we see the value of including MI in speech recognition applications. The 0 values indicate a complete lack of variation among the dialects in that cluster for that vowel pair. That is, if there is complete predictability for one of the words, then knowing about the other cannot improve our predictions of the first. Aside from these cases of 0's, including MI would always improve performance. This finding is in keeping with what has been shown for MI as applied to handprinting recognition [4].

Finally, in an effort to determine the extent to which a subset of the phonemic features determines the dialect, and the remaining features, we automatically classified the varieties into clusters using a Nearest Neighbor classifier algorithm and a leave-one-out design for partitioning the samples [22]. We computed the error rate of misclassifying the dialect cluster of the variety, given the remaining varieties of that dialect cluster, at various threshold levels. With K=3, the dialects were classified with 6 errors, an error rate of 0.10.

## 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 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 phonological context. Given access to appropriate data, perhaps from [10-12], 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 does 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, cell-phone 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 dependences in the pronunciation of speakers of 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 [28], 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.

## 8 References

1. Fetzer, A. *Recontextualizing context: Grammaticality meets appropriateness.* 2004. Benjamins: Philadelphia.
2. Giunchiglia, F. and P. Bouquet, Introduction to contextual reasoning. An Artificial Intelligence Perspective, in *Perspectives on Cognitive Science 3*, B. Kokinov, Ed. 1997, NBU Press: Sofia, Bulgaria.
3. Sarkar, P. and G. Nagy, Style consistent classification of isogenous patterns. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2005. 27(1):88-98.
4. Veeramachaneni, S. and G. Nagy, Style context with second order statistics. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2005. 27(1):14-22.
5. Carver, C.M. *American Regional Dialects: A Word Geography*. 1987. University of Michigan Press: Ann Arbor.
6. Labov, W., S. Ash, and C. Boberg. *Atlas of North American English.* 2005. Mouton de Gruyter: Paris.
7. Hughes, A. and P. Trudgill. *English Accents and Dialects: An Introduction to Social and Regional Varieties of British English.* 1987. Edward Arnold: London.
8. Trudgill, P. *The Dialects of England.* 1999. Blackwell: London.
9. Nerbonne, J. and P. Kleiweg, Lexical distance in LAMSAS. *Computers and the Humanities*, 2003. 37(3):339-57.

10. Gooskens, C. and W. Heeringa, Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. *Language Variation and Change*, 2004. 16(3):189-207.

11. Cheng, C.-C., Measuring Relationship among Dialects: DOC [Dictionary on computer] and Related Resources. *Computational Linguistics and Chinese Language Processing*, 1997. 2(1):41-72.

12. Heeringa, W. and A. Braun, The Use of the Almeida-Braun System in the Measurement of Dutch Dialect Distances. *Computers and the Humanities*, 2003. 37(3):257–71.

13. Heeringa, W., *Measuring dialect pronunciation differences using Levenshtein distance*. 2004, University of Groningen: Groningen.

14. Heggarty, P.A. *Measured Language: From First Principles to New Techniques for Putting Numbers on Language Similarity*. in prep. Blackwell: Oxford.

15. Schneider, E.W., et al., eds. *A Handbook of Varieties of English: A Multimedia Reference Tool*. 2005, Mouton de Gruyter: Berlin, New York.

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

17. Wells, J.C., ed. *Accents of English*. 1982, Cambridge University Press: Cambridge.

18. Kaufman, L. and P.J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. 1980. Wiley: Hoboken, NJ.

19. Day, W.H.E. and H. Edelsbrunner, Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification*, 1984. 1(1):7-24.

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

21. Theodoridis, S. and K. Koutroumbas. *Pattern Recognition*. 1999. Academic: NY.

22. Duda, R.O., P.E. Hart, and D.G. Stork. *Pattern Classification*. 2001. Wiley-Interscience: Hoboken, NJ.

23. Topchy, A., et al. Adaptive Clustering Ensembles. *Proc. ICPR*. 2004. Cambridge.

24. Jain, A.K., et al. Landscape of Clustering Algorithms. *Proc. ICPR*. 2004. Cambridge.

25. Redner, R.A. and H.F. Walker, Mixture densities, maximum likelihood, and the EM algorithm. *SIAM Review*, 1984. 26(2):195-235.

26. Topchy, A., A.K. Jain, and W. Punch. A Mixture Model for Clustering Ensembles. in *Proc. SIAM International Conference on Data Mining (SDM04)*. 2004. Florida.

27. Foulkes, P., Current trends in British sociophonetics. *Univ. of PA Working Papers in Linguistics: A Selection of Papers from NWAV 30*, 2002. 8(3):75-86.

28. Rabiner, L.R. and B.H. Juang. *Fundamentals of Speech Recognition*. 1993. Prenctice Hall: Englewood Cliffs, NJ.