A Quantitative Categorization of Phonemic Dialect Features in Context

Naomi Nagy — University of New Hampshire — ngn@unh.edu

Xiaoli Zhang — Rensselaer Polytechnic Institute — zhangxl@rpi.edu

George Nagy — Rensselaer Polytechnic Institute — nagy@rpi.edu

Edgar W. Schneider — Regensburg University — edgar.schneider@sprachlit.uni-regensburg.de

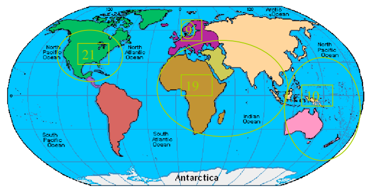
Summary

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 co-occurring 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.

Next steps

- Test these methods at all levels of the continuum from idiolect to language, using many idiolects from each dialect
- Predict, for a partially unanalyzed dialect, what features it will exhibit (based on knowledge of some subset of features that it does exhibit)
- Apply to speaker identification
  - stochastic description of a speaker's full dialect
  - base on a sample containing a subset of phonemes
- Automated speech recognition
  - accuracy could be raised by exploiting the consistency and the statistical dependencies in the pronunciation of speakers of a given dialect cluster

Data Organization



(www.jayzeebear.com/ map/world.gif)

The list of vowel features builds on the **lexical sets** devised by J.C. Wells, a system of distinct vowel types identified by key words (e.g. KIT for the vowel in *this* and *ridge*; DRESS for the vowel in *bet* and *said*).

Possible variants of the vowel of KIT:

(1) canonical or basic high front [ɪ]
(2) raised and fronted [i] (as in *seed*)
(3) centralized [ə] (as in *cup*)
(4) with an offglide, e.g. [iə/ɪə]

Feature type	# features	#variants
Vowel	28	121
Vowel merger	4	4
Consonant	32	38
Prosodic	5	5
TOTAL	69	168

- Each element w_{ij} corresponds to a variant of a phonological feature for variety V_i
- 69 phonological features F_i,
  - 2-7 variants (possible values) per feature
- Each binary feature vector $\mathbf{w}_i$ has 168 elements (of which 13 are shown here).

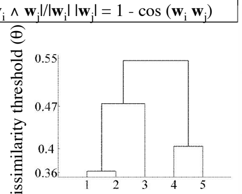
(W) Binary features (F_i) for two vowels in 13 dialects (V_i)

Variety	KIT central	KIT raised	KIT basic	DRESS close	DRESS open
Orkney & Shetland	0	0	1	0	1
North of England	1	0	0	0	1
East Anglia	1	0	0	0	1
Philadelphia	1	0	0	0	1
Newfoundland	0	0	1	0	1
Cajun English	1	0	0	0	1
Jamaican Creole	0	1	0	0	1
Tobago Basilect	1	0	0	0	1
Australian Creoles	1	0	0	1	0
Tok Pisin	0	1	0	1	0
Fijian English	0	0	1	0	1
Nigerian Pidgin	1	0	0	0	1
Cape Flats English	0	0	1	1	0
TOTAL	**7**	**2**	**4**	**3**	**10**

Methods: Clustering

- Complete Link Algorithm to create clusters
- Clusters are merged when the maximum dissimilarity between a variety in one cluster and a variety in the other cluster is $< \theta$.

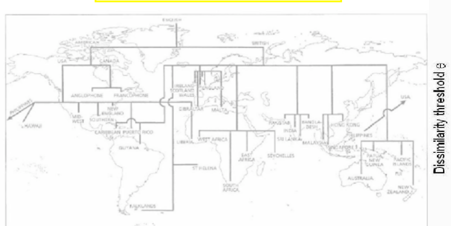
Dissimilarity ρ_{ij} between 2 varieties = $1 - |\mathbf{w}_i \wedge \mathbf{w}_j| / |\mathbf{w}_i|\,|\mathbf{w}_j| = 1 - \cos(\mathbf{w}_i\ \mathbf{w}_j)$



Varieties	1	2	3	4	5
1	0	.36	.40	.46	.50
2		0	.47	.44	.40
3			0	.50	.55
4				0	.55
5					0

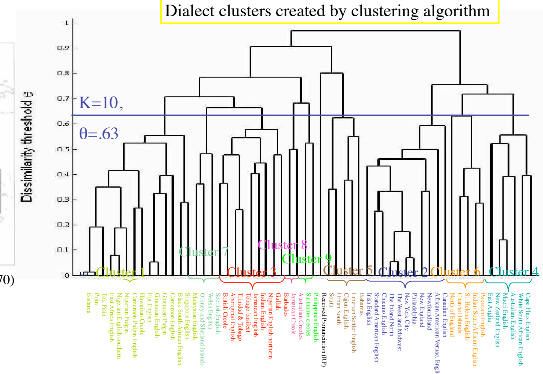
Results: Clustering

Traditional family tree model



(Crystal 2003:70)

Dialect clusters created by clustering algorithm



Methods: Mutual Information (MI)

- The amount of context = the average MI between pairs of features.
- MI is based on the marginal and joint probabilities of the features within a cluster.
- MI = the relative entropy between the two distributions: MI indicates how much each distribution reveals about the other.

Calculation of 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 two features in 13 dialects

Calculation of Mutual Information			KIT basic	KIT raised	KIT central
			.54	.16	.30
DRESS	close	.24	.08	.08	.08
DRESS	open	.31	.76	.08	.22

6 individual components of MI

$I(x_i, y_j) =$	-.06	.08	.01
$(\Sigma = .05)$	.08	-.05	-.01

$I_{DRESS,KIT} = 0.05 < H(x) = 0.54 < \log_2 2 = 1.00$; $H(y) = 1.41 < \log_2 3 = 1.59$

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

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

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

Results: Mutual Information

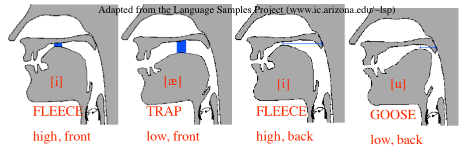
- There is a degree of MI across *every* pair—
- any word recognition application would be improved by including MI in its calculations.
- (There are *no* cases of completely independent variation)

F1 \ F2	KIT (Lax)	DRESS (Lax)	FOOT (Lax)	THOUGHT (Lax)	FLEECE (Tense)	FACE (Tense)	GOAT (Tense)	GOOSE (Tense)
KIT	2.00	0.41	0.58	0.33	0.52	0.61	0.69	0.51
DRESS		1.48	0.13	0.30	0.24	0.30	.040	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

=highest values

=auto-comparisons

Adapted from the Language Samples Project (www.ic.arizona.edu/~lsp)



Results: Clustering and Mutual Information

MI for 4 tense and 4 lax vowels	Cluster 1 (K = 10, θ = 0.63)	2	3	4	5	6
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, 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
...						

Highlighted cells show 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).

"0" = no variation within that cluster for that vowel pair: 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, MI would always improve performance of ASR.