Dialectal Clustering Using Exemplar-based Models of Phonotactics

[name]
[address1]
[address2]
[address3]
[email]

[name]
[address1]
[address2]
[address3]
[email]

[name]
[address1]
[address2]
[address3]
[email]

Extended Abstract

1 Introduction

L2 speakers of a language show systematic regularities in handling their new language's unfamiliar phonotactics. Some of these regularities are produced by other speakers of their L1, while some regularities are unique to the individual. The discovery of these group traits in phonotactics allow us to define dialectal models.

Computational dialectal modeling has been pioneered by John Nerbonne (2005, 2006) Nerbonne's dialectal work, however, primarily utilizes single token comparisons,and Levenshtein feature distance. Nerbonne has also shown an interest in phonotactics, but he has never united his phonotactic and dialect work. (Nerbonne, 2004; Sang & Nerbonne, 2000)

The method we adopt is constraint-based, and the rankings of the constraints represent the frequency of exposure to each bigraph, internally these are represented as markedness constraints, such as *Phoneme1Phoneme2, (e.g. *[kp] is a highly ranked constraint, while *[pa] is lowly ranked).

The corpus we used was the George Mason Speech Accent Archive, a database of 147 English L1 and 377 nonnative speakers of English reading the same 4-sentence segment. (Weinberger, 2006) This database can be queried by native language, amount of English exposure, geography, age, sex, and numerous other biographical and linguistic variables.

One Spanish L1 speaker read this text as the following:

[plˌis kol ɛstɛl̼a æks hɛɹ tu briˇŋ+ d̼is θiˇŋ+s wiθ hɛɹ frãm d̼ɛ stoɹ siks ɛspuˇns oɹ fɹɛʃ ɛsno pis faɪf θik ɛslæf of blu ʧis ãm mebi ɛ snækˈ foɹ hɛɹ broðɛɹ bapˈ wi also nid ɛˇ smol plˌæstikˈ ɛˇsnek æˇn ʌ bikˈ toɪ frak foɹ də kis ʃi kæn ɛskupˈ ðis θɪˇŋ+s iˇntu tɹi rɛt bæks æˇŋ+ wi wil go mitˈ hɛɹ wɛˇzde æˇd̼ə tɹeˇn ɛsteʃəˇn]

Across the majority of Spanish speakers, [ɛ] or [ə] was epenthesized before [s] onsets, regressive assimilation of the [+nasal] feature occurred effecting vowels, and [ð] was often produced as [d]. Documenting these regularities can provide us with a model of Spanish pronunciation, while the ability to learn these (and other) regularities across a corpus allows us to discover many divergent subgroups.

Our Gold Standard pronunciation comes from the Webster's Pocket Lexicon, as incorporated into the Hoosier Mental Lexicon, a database of 19323 words, along with their broad phonetic transcription, metrical structure, and frequency. (Nusbaum et. al., 1994)

For each speaker in this database, we compute markedness constraints of bigraphs, and sum over all the bigraphs in the word.

$$\text{Word Markedness} = \frac{\sum_{i=0}^{n-1}((log_2\ e(i)*e(i+1))/log_2\ o(i,i+1))}{n}$$

This algorithm is based on an exemplar-based Optimality Theoretic model of phonotactic learning introduced by Mosier (2003). For every bigrams in the text, sum the log of the expected probability of the bigram divided by the log of the observed frequency. This number is then normalized by dividing by the length of the word. Mosier did not include positional markedness, but we have three: beginning, middle, and end.

We graphing the results utilizing overall per-speaker markedness values, speaker inventory, and gold standard-divergent word-markedness values, showing that speakers of an L2 coming from the same L1 show a number of group traits, and that these group traits can be learned algorithmically and exploited for dialect identification.

A useful application of this accent difference algorithm is the quantitative assessment of L2 pronunciation during language learning. The most important theoretical application is the quantitative measurement of dialectal and cross-linguistic distance beyond single token-based feature differences.

References

Billerey-Mosier, Roger. Exemplar-based phonotactic learning. SWOT 2003. April 05, 2003.

Nerbonne, John. Linguisticss in Aggregate Comparison. Literary and Linguistic Computing 21(4), 2006. (J.Nerbonne & W.Kretzschmar, Jr. (eds.) Progress in Dialectometry: Toward Explanation (SUBMITTED)

Nerbonne, John and Peter Kleiweg. Toward a Dialectological Yardstick. Accepted (5/2005) to appear in: Journal of Quantitative Linguistics.

Nerbonne, John and Ivilin Stoianov. Learning Phonotactics with Simple Processors. In: Dicky Gilbers, Maartje Schreuder and Nienke Knevel (eds.)On the Boundaries of Phonology and Phonetics CLCG: Groningen, 2004, pp.89-121.

Nusbaum, H.C., D.B. Pisoni, and C.K. Davis. 1984. Sizing up the Hoosier Mental Lexicon: Measuring the familiarity of 20,000 words. Research on Speech Perception Progress Report No. 10, 357-376. Bloomington, IN: Speech Research Laboratory, Indiana University

Sang, Erik Tjong Kim and John Nerbonne. Learning the Logic of Simple Phonotactics. In: James Cussens and Saso Dzeroski (eds.) Learning Language in Logic. Springer Lecture Notes in Artificial Intelligence. New York and Berlin: Springer, 2000,

Weinberger, Steven H. Speech Accent Archive. George Mason University. April 10, 2006. Retrieved from http://accent.gmu.edu