Hierarchical statistical inference and lexical diffusion of sound change

Vsevolod Kapatsinski

University of Oregon

LSA Annual Meeting
Boston, MA
January 3-6, 2013
Goals of this talk

• Discuss the relationship between usage-based linguistic theory and mixed-effects statistical models (particularly partial pooling, Zipf’s law and lexical diffusion)
  – Implications for data analysis
  – Implications for theory
Two kinds of change in Usage-based Phonology

• Articulatorily-motivated sound change
  – Tempered by avoidance of misperception (Lindblom 1990)
  – Starts in high-frequency words
    • E.g., word-final t/d deletion (Bybee 2002), memory/mammary (Hooper 1976)
    • But what about analogy to reduced words?

• Analogical change
  – Driven by pressure to be like similar items (analogy) and imperfect learning
  – Starts with low-frequency words
    • e.g., irregular past tenses are all in high-frequency verbs
Implementing the opposing mechanisms

• Reduction in use: Every time a word is used, it reduces
  – Words are assumed to be units of articulatory planning and execution (Bybee 2001, 2002, Kapatsinski 2010)
  – Or at least used in a reduction-favoring context (Antić 2012, Bybee 2002)

  – “word-specific phonetics”
Is sound change in the phoneme or in the word?

• It’s in both.
• Phonemes and words can be associated with rates of reduction (cf. Phillips 2001).
• Ascribing blame for reduction to phoneme vs. word is a process of hierarchical statistical inference.
• We implement this using lme4 in R (Baayen et al. 2008)
Predicting reduction

\[ p(\text{reduction}) \sim \beta_0 + \beta_1 \times \text{word} \]

- \( \beta_0 \) = overall probability of reduction (for this gesture/phoneme)
- \( \beta_1 \) = adjustment associated with individual word
A learning problem: Zipf (1949)

• Zipf’s law:

• Consequence of a rich-get-richer effect (positive feedback) (Barabasi 2002)

• In a corpus of any size, most words occur only once (Baayen 2001)

→ For most words, we cannot estimate a word-specific reduction coefficient with any certainty
The standard solution (mixed-effects): Partial pooling / Word as a random effect
(Gelman & Hill 2007: 252-259)

• Assume that the coefficients associated with individual words come from a certain distribution

• Allow coefficients to constrain each other: Coefficients that are too far off from the center of the distribution are pulled in unless supported by much data

• Weighted average of the mean of the tokens of the word ($\overline{y}_j$) and the mean of the tokens of the phoneme/gesture ($\overline{y}_{all}$)

$$\frac{n_j \overline{y}_j + 1 \overline{y}_{all}}{\sigma_y^2 + \frac{1}{\sigma_\alpha^2}}$$

(Gelman & Hill 2007: 253)

Note: $n_j$ is the number of tokens of the word: low-frequency word averages are less reliable and get pulled in to the mean for the phoneme/gesture
Note: this does not assume a normal distribution
The child as a mixed effects model

• Recall the theory:
  – Every time you use a word, its reduction probability is incremented
  – When children acquire language, they acquire word-specific and phoneme/gesture-specific phonetics
    • They do not try to recover the coefficient associated with word frequency
    • Zipf’s law applies
    • This is equivalent to saying the child is modeling reduction as an overall rate for a phoneme with a random deflection for each word
 Evidence: Erker & Guy (2012)

- Pronoun use with Spanish verbs
- Grammatical effects are augmented in high-frequency words
- As you would expect if within-word coefficients for grammatical predictors are calculated by the learner
Partial pooling and word frequency

• Lexical diffusion predicts that \( y \) systematically varies with \( n \)
  – Low-frequency words are predicted not to behave like high-frequency words

• Yet, partial pooling assumes that we can estimate reduction probabilities associated with low-frequency words on the basis of those associated with high-frequency words

→ Mixed-effects models with word as a random factor will systematically underestimate word frequency effects and attribute effects due to frequency to individual words

Prediction: At late stages of an articulatorily-driven change, exceptionally conservative words are likely to have intermediate frequency of use. When high-frequency words come to be associated with reduced variant of the phoneme, low-frequency words will be pulled in.
The U-shaped frequency effect
Parameters

• Number of distinct words potentially affected by the change / containing the phoneme

• Size of the reductive effect of frequency
  – How much is the likelihood of reduction incremented for every additional token of use

• Word frequency distribution
  – How much do words potentially affected by the change differ in frequency
  – Implemented by different limits on uniform distributions on log(freq) continuum and exponents of the Zipf curve
Predictions

\[ \frac{n_j}{\sigma_y^2 y_j + \frac{1}{\sigma_\alpha^2 y_{all}}} \]  (Gelman & Hill 2007: 253)

Small differences between words
(Due to differences in frequency or weak frequency effect on reduction)

→ Low-frequency words more likely to behave like high-frequency words

A lot of super-low-frequency words → more uniformity

No effect of lexicon size / type frequency beyond the fact that larger lexicons contain more super-low-frequency words (Baayen 2001)
Conclusion

• Zipf’s law makes it impossible to reliably estimate the behavior of a huge proportion of super-low-frequency words in the lexicon (hapax legomena).

• Mixed-effects models / partial pooling provides a solution to this issue by letting estimates be made on the basis of the behaviors of higher-frequency words.

• Word frequency effects will be systematically underestimated by both analysts and learners.

• Lexical diffusion of articulatorily-motivated sound changes is predicted to affect high-frequency words, then low-frequency words, with some mid-frequency words remaining exceptionally non-reduced.

• The low-frequency words are especially likely to come to behave like high-frequency words in later stages of the change if:
  – Learners are conservative in estimating word-specific reduction probabilities, not assuming that something cannot be reduced after hearing it once.
  – The frequency effect on reduction is weak.
  – Words containing the phoneme do not differ much in frequency.
References


