G2P Conversion of Proper Names Using Word Origin Information
Sonjia Waxmonsky and Sravana Reddy, The University of Chicago

Does knowing the language of origin help predict a name’s pronunciation?

SH in Schoenberg (German)
S K in Schiavone (Italian)
J in Judd (English)
Y in Jung (German)
I in Juarez (Spanish)

Language-Aware Pronunciations
Train multiple language-specific grapheme-to-phoneme models to improve probabilistic G2P conversion

Step 1
- For words of unknown origin, train a word origin model to predict
  \[ P(\text{Lang} | \text{Word}) = \frac{\sum_{L} P(\text{Lang} | \text{Word}) P(\text{Pronun} | \text{Word}, \text{Lang})}{\sum_{L} P(\text{Lang} | \text{Word})} \]

Step 2
- For each Lang, make training dictionary containing all words where
  \[ P(\text{Lang} | \text{Word}) > 0.7 \]
- Train language-specific G2P model

Step 3
- Method A: Weight G2P output by word origin probability
- Method B: Smooth results against a language-independent model with a factor \( \alpha \). In our experiments, \( \alpha = 0.5 \) by tuning on a development set.

Data
- Names were queried against the CMU Dictionary. Those with known pronunciations are retained, giving a set of 46k names.
- 6% of names were hand-annotated for language of origin.
- 12 most frequent languages plus “Other” class form a set of languages \( L \).
- Annotated data is available at people.cs.uchicago.edu/~wax/wordorigin/

Results
- The G2P algorithm used for all experiments is Sequitur\(^1\) with 4-grams
- Data is split 80/10/10 into train/dev/test
- Baseline: fully language-independent G2P

Method A is slightly worse than the baseline because of data sparsity in language-specific G2P models.
Method B is better than the baseline, because it captures language-specific pronunciations as well as Americanization patterns.

Grapheme-to-Phoneme

Graphemes
- schoenberg

Stochastic Model

Phonemes
- SH OW N B ER G