The Acquisition of Vowel Harmony from Simple Local Statistics

Spencer Caplan  Jordan Kodner

Vowel Harmony

- System-wide vowel alternation patterns in languages across the world
- Affects roots and affixes, and languages may have multiple processes
- Vowels are either neutral or harmonizing. Harmonizing vowels are partitioned into sets.
- Generally, words contain vowels from only one harmonizing set

Early Acquisition

- Infants as young as seven months are sensitive to vowel harmony alternations in acoustic input in preferential listening over continuous unsegmented speech [4]
- A computational model should:
  - Connect raw input to phonological theory
  - Require little data and function over unsegmented speech (rather than individual words from a wordlist)
  - Leverage plausible cognitive tools (online processing, simple calculations, innate ability to differentiate consonants and vowels) [1]

Distributional Hypothesis

1. In non-harmonizing languages (eg English), no particular vowels should be more or less likely to follow one another (near-uniform co-occurrence distribution)
2. In harmonizing languages (eg Turkish, Finnish), we expect strongly non-uniform co-occurrence distributions, since vowel co-occurrence is partitioned by the phonology

Model Implementation

1. Over word list OR unsegmented input,
2. Tabulate vowel co-occurrence matrix counting adjacent vowels ignoring consonants
3. Convert to normalized conditional probabilities
4. Remove neutral vowels with normP consistently below threshold proportional to vowel set cardinality
5. Find featural partition (eg online k-means clustering, k=2)
6. Collapse over features and repeat.

Distributional Cues

- Heatmaps showing normP for each vowel pair in Finnish, Turkish, and English
- Red indicates low normP (unlikely to co-occur) and Green is high normP
- Self-normP is omitted

Results

Language | Primary Harmony | Secondary Harmony | Neutral Vowels | Harmony Found
--- | --- | --- | --- | ---
Hungarian | yes | yes | 2 | Primary only
Turkish | yes | yes | (2ary only) | yes
Finnish | yes | - | 2 | yes
Warlpiri | yes | - | 1 | yes
Uyghur | yes | - | - | yes
Estonian | remnant | - | - | finds remnant
German | - | - | - | no
English | - | - | - | no

Empirical Questions

- Is primary harmony in fact acquired first?
- Are children sensitive to more complex harmony processes early on?
- Must harmony function over a single phonological feature?
- How do learners differentiate productive harmony (eg Finnish, Turkish) from non-productive (eg Estonian, Uzbek)? [cf 5]
- How helpful is harmony in word segmentation tasks?
- What minimum signal-to-noise ratio is necessary? Harmony exceptions (eg common in Turkish) and average word length in infant-directed-speech affect this.

Acknowledgements

We thank Charles Yang, Mitch Marcus, Ryan Budnick, the University of Pennsylvania's 2016 LORELEI research team, and the Penn Language Development and Language Processing Lab. This research was funded by the DARPA LORELEI program under Agreement No. HR0011-15-2-0023.

Selected References


Code and Contact

https://github.com/scaplan/VowelHarmonyAcquisition
{scaplan, jkodner}@sas.upenn.edu