Discovering Phonesthemes with Sparse Regularization

Nelson F. Liu♠♡, Gina-Anne Levow♡, Noah A. Smith♠

♠Paul G. Allen School of Computer Science & Engineering; ♡Department of Linguistics, University of Washington

UWNLP

Abstract

- Goal: Extracting non-arbitrary form-meaning representations (phonesthemes) from word vectors.
- Treat the problem as one of feature selection for a model trained to predict word vectors from subword features.
- Many model-predicted phonesthemes overlap with those proposed in the linguistics literature; human judgments provide additional validation.

Phonesthemes

English: apple

Chinese: ping guo

French: pomme

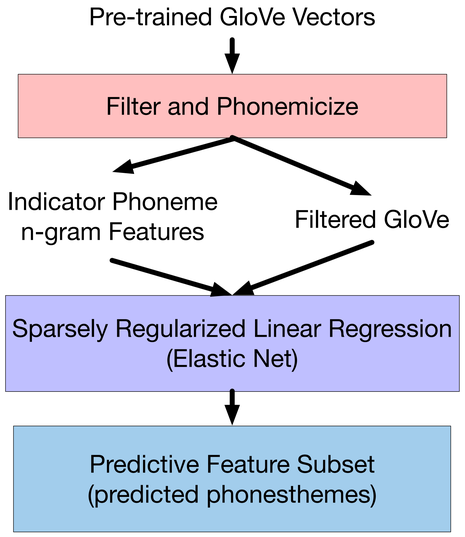
Arabic: tafaha

No link between what an object is and how it's represented in language (de Saussure, 1916).

Phonesthemes: non-compositional, submorphemic phonetic units that consistently occur in words with similar meanings

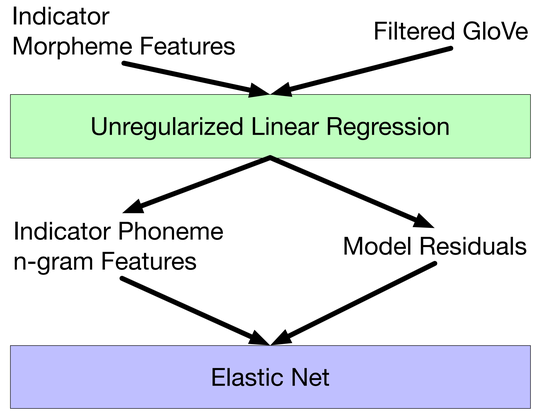
- "gl-" : "glow", "glint", "gloss"
- "sn-" : "snarl", "sniff", "snooty"

Phonestheme Identification as Sparse Feature Selection



Reducing the Effect of Morphemes

- **Intuition: remove vector components predictable from morpheme-level information**



Model-Predicted Phonesthemes

	Character Sequence	Model Example Words
† *	*sn-*	*sneaks*, *snubs*, *sniffs*
*	*sc-/sk-*	*screwing*, *squelched*, *scurry*
*	*cr-*	*crunched*, *cringed*, *crummy*
*	*sp-*	*spiffy*, *splendidly*, *spunky*
	br-	*brags*, *brouhaha*, *brutish*
*	*gr-*	*griping*, *grumbles*, *grandly*
*	*tr-*	*tryst*, *trounce*, *truism*
*	*st-*	*stupendous*, *startlingly*, *stunner*
† *	*bl-*	*blase*, *blithely*, *blankly*
*	*fl-*	*flaunted*, *flowered*, *fluff*
† *	*gl-*	*glossed*, *gleam*, *glamor*
*	*sl-*	*slouch*, *slogged*, *slime*
† *	*dr-*	*droll*, *dreamer*, *drifter*
† *	*sw-*	*swoon*, *swoops*, *swipes*
	wi-	*wimpy*, *willy*, *wince*

Character Sequence	Model Example Words
ca-	*candied*, *caffeinated*, *cataclysm*
pa-	*pantry*, *pathogen*, *pancake*
sy-/si-	*syllable*, *simulators*, *synchronize*
fr-	*froth*, *frock*, *freaks*
ma-	*mallet*, *masts*, *manor*
pe-	*pendant*, *pelt*, *petulant*
me-	*meld*, *meditate*, *memorized*
mu-	*mumbled*, *mummies*, *mutter*
* *cl-*	*clumsily*, *clunky*, *claustrophobic*
se-/ce	*sensuous*, *celibate*, *celebrants*
ob-	*obliterate*, *abridged*, *obliquely*
ba-/bo-	*barbarous*, *bogs*, *barbers*
pl-	*pled*, *pliable*, *platoons*
co-	*corset*, *coroners*, *corduroy*
fe-	*fairest*, *fender*, *feds*

Table 1: The 30 model-predicted phonesthemes with the highest absolute model weight (1-15 on left, 16-30 on right). * indicates a phonestheme identified by Hutchins (1998). † indicates a phonestheme with statistical support from Otis and Sagi (2008).

Human Judgments

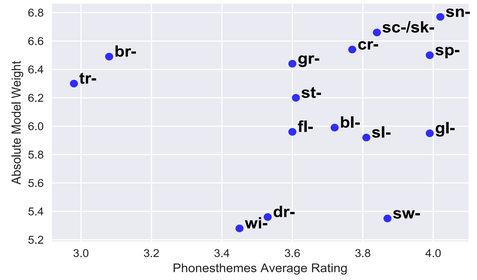


Figure 1: On average, model-predicted phonesthemes were rated 0.58 points higher than unselected phonesthemes (3.66 versus 3.08).

References

- E. Dario Gutierrez, Roger Levy, and Benjamin Bergen. 2016. Finding non-arbitrary form-meaning systematicity using string-metric learning for kernel regression. In *Proc. of ACL*.
- Sharon Suzanne Hutchins. 1998. *The psychological reality, variability, and compositionality of English phonesthemes*. Ph.D. thesis, Emory University.
- Katya Otis and Eyal Sagi. 2008. Phonaesthemes: A corpus-based analysis. In *Proc. of CogSci.*
- Ferdinand de Saussure. 1916. *Course in General Linguistics.*