Speech production in the discriminative lexicon

Word and Paradigm Morphology with Linear Discriminative Learning

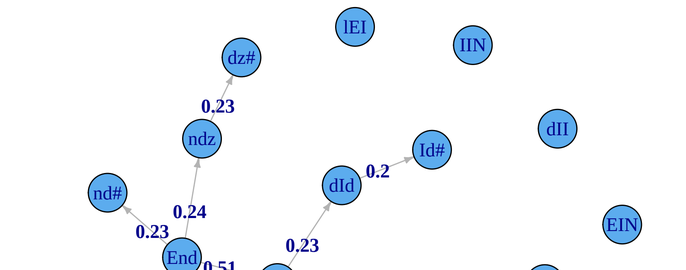
R. Harald Baayen, Yu-Ying Chuang, and James P. Blevins
Quantitative Linguistics Lab, Department of Linguistics, Eberhard Karls University Tübingen
harald.baayen,yu-ying.chuang@uni-tuebingen.de, jpb39@cam.ac.uk

goals

1. develop algorithms for comprehension and production that do not require theoretical constructs such as morphemes, allomorphs, exponents, stems, and inflectional classes, i.e., provide an implementation of word and paradigm morphology
2. show that 'morphological' effects in the experimental literature follow straightforwardly

background: mathematics of linear mappings

$$\boldsymbol{C} = \begin{pmatrix} 1 & 2 \\ -2 & -2 \\ -2 & 1 \end{pmatrix} \qquad \boldsymbol{S} = \begin{pmatrix} 2 & -4 \\ -4 & 4 \\ -4 & -2 \end{pmatrix}$$



representations

form : vectors encoding which triphones are present in a word's form (1: present, 0: absent)

meaning : real-valued vectors (LSA, HiDEx, word2vec, or NDL); for inflected words, the semantic vectors of base and affixal function are summed

algorithms

comprehension

- map form vector on semantic vector with $\boldsymbol{F}$
- select closest neighbor semantic vector as top candidate

production

- map semantic vector onto form vector with $\boldsymbol{G}$
- select the form with best supported path in the production graph as top candidate

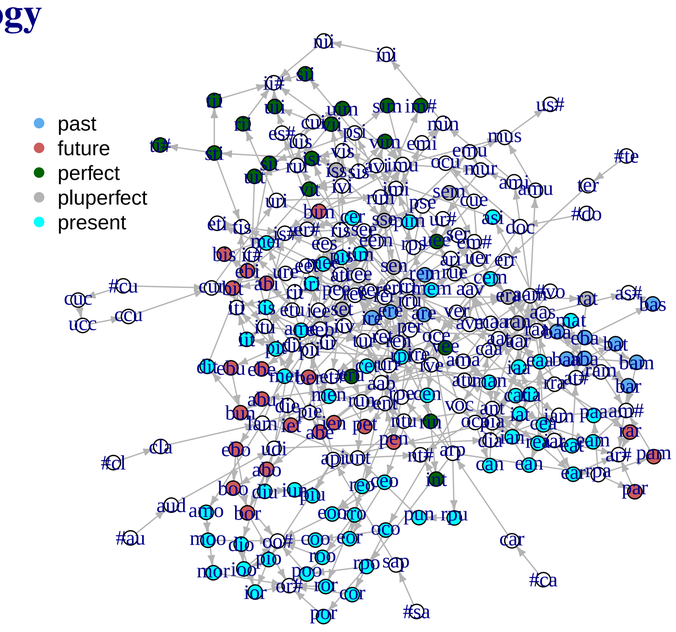
thinned directed graph path with paths for inflected forms of English blend*, given the semantic vector of* blending *as input; weights on edges represent the activations of the corresponding end vertices*

accuracy

language	semantic vectors	dataset	accuracy comprehension	accuracy production
English	TASA	11480 words	dual: 8954/11480	10562/11480
Latin	simulation	$14 \times 6 \times 8$ verb forms	672/672	670/672
German	simulation	76×8 nouns	601/608	608/608
Dutch	simulation	final devoicing words	3381/3392	3387/3392
Dutch	simulation	final devoicing nonwords		171/192
Polish	simulation	$2 \times 7 \times 66$ noun forms	918/924	923/924
Estonian	simulation	$2 \times 10 \times 28$ noun forms	560/560	560/560
Hebrew	TNK	4153 QAL forms, 437 roots	4054/4153	4152/4153

simulated semantic vectors are constructed such that forms sharing content or grammatical lexomes are more correlated compared to forms that do not share any lexomes

brain topology



for the Latin dataset, self-organization of triphones in 2-D with the graphopt algorithm reveals areas where triphones predominantly support specific tenses: topographical 'morphemic' effects (in, e.g., fMRI) may thus have a non-morphemic source

speech errors

Latin *curriaaris* for *curraaris* by analogy to *audiaaris*; Dutch *ruizen* for *ruisen*, Estonian *kaldas* (in sg) for *kallas* (nom sg); Hebrew *yipgoxna* for *tipgoxna* (first person singular instead of second person singular); English *mouths* with voiceless instead of voiced th

orthographic form	content lexome	phonetic form (DISC)	grammatical lexome
slice	SLICE	sl2s	PRESENT
sliced	SLICE	sl2st	PAST
slicing	SLICE	sl2sIN	GERUND
nice	NICE	n2s	
nicely	NICE	n2slI	ADVERB
price	PRICE	pr2s	PRESENT
pricely	PRICE	pr2slI	ADVERB
thin	THIN	TIn	PRESENT
thinned	THIN	TInd	PAST
thinning	THIN	TInIN	GERUND
thinly	THIN	TInlI	ADVERB
genuine	GENUINE	JEnjUIn	
genuinely	GENUINE	JEnjUInlI	ADVERB

the speech error slicely thinned *arises when initially the semantic vector of* SLICE *is integrated with that of* ADVERB *and mapped onto the triphones, and subsequently the semantic vector of* THIN *is integrated with that of* PAST*, and mapped onto the triphones (muddled thinking for speaking)*

boundary effects

boundary effects (delayed inter-keystroke intervals at morpheme boundaries; longer segment durations in the speech signal) arise due to weakly supported edges in the production graph, typically where the graph branches (see example of *blending*); for English, edge weights predict pertinent segment durations in the Buckeye corpus

machine learning ethics

open source code, straightforward algorithms, no frankenalgos, no risk of black swans

references

Baayen, R. H., Chuang, Y. Y., and Blevins, J. P. (2018). Inflectional morphology with linear mappings. *The Mental Lexicon* **13.2**.

Baayen, R. H., Chuang, Y. Y., Shafaei-Bajestan, E., and Blevins, J. (2018). The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de)composition but in linear discriminative learning. Manuscript under review for *Complexity*.

Baayen, R. H., Chuang, Y. Y., and Heitmeier, M. (2018). WpmWithLdl (version 1.0). R-package available at http://www.sfs.uni-tuebingen.de/~hbaayen/publications/WpmWithLdl_1.0.tar.gz

acknowledgements

this research was funded in part by the European Research Council (grant # 742545 (WIDE) to the first author); we are indebted to Kaidi Lõo for the Estonian dataset, to Paula Orzechowska for the Polish dataset, Maria Heitmeier for the German dataset, and to Dirk Roorda and Martijn Naaijer for the Hebrew dataset.