The relationship between lexical frequency, compositionality, and phonological reduction in English compounds

Forrest Davis and Abigail C Cohn

Cornell University

94th Annual Meeting of the Linguistic Society of America, January 2020

Research Questions

- Are more opaque compounds (*cupboard*) phonologically different from more transparent compounds (*blueberry*)?
- Are effects of compositionality distinct from effects of lexical frequency and degree of conventionalization?

Introduction

reduced	not reduced
cúpboard	sóngbìrd

- **Compositionality** Degree that the meaning of a compound is the sum of its parts (e.g. *humbug* vs. *blueberry*)

less compositional	more compositional
opaque	**transparent**

- Is this relationship between phonological reduction and **compositionality** (Libben and Jarema, 2006) robust? Is it distinct from those of **lexical frequency** (Jurafsky et al., 2001; Bell et al., 2009)?

Data

- Buckeye Corpus (Pitt et al., 2007)
- Used 21 most frequent bisyllabic nominal compounds orthographically represented with no space (e.g. *roommate*, *airline*, *freshman*, *football*)

Hypothesis

More **opaque** compounds are more phonologically **reduced** than transparent ones

Measure of Compositionality

- Goal to establish a gradient measure of compositionality (cf. Libben and Jarema, 2006)
- Survey of 24 native American English speakers using a 7 point Likert Scale

cupboard, stalemate		*blueberry, doorbell*
1	4	7
very opaque	**neither opaque nor transparent**	**very transparent**

Measures of Lexical Frequency

- Frequency for **compound** and its **constituents** (e.g. *homework*: 6069, *home*: 196061, *man*: 216061)
- Counts with add-one smoothing Corpus of Contemporary American English (COCA) (Davies, 2008)
- **Pointwise mutual information (PMI)** calculated (e.g. *homework*: 4.16, *freshman*: 15.95); correlated with conventionalization (Evert, 2008; Ramsich et al., 2010)

$$\text{PMI}(xy) \equiv log\frac{p(xy)}{p(x)p(y)}$$

Main Results

- Rating and PMI (degree of conventionalization) are significant predictors of final rime duration
- The less compositional a compound the shorter its final rime

Reduction Results

- Mean duration of the final rime shorter when ratings are low (**opaque**) than expected given rime duration in non-compounds (e.g. *ware* in *software* is half the duration of *where*)
- Rating and PMI statistically significant predictors ($p < 0.005$)
- Modifier frequency and duration in monosyllabic compounds also statistically significant ($p < 0.05$)

Selected References

Alan Bell, Jason M Brenier, Michelle Gregory, Cynthia Girand, and Dan Jurafsky. 2009. Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language, 60*.

Daniel Jurafsky, Alan Bell, Michelle Gregory, and William D Raymond. 2001. Probabilistic relations between words: Evidence from reduction in lexical production. *Typological Studies in Language, 45*.

Gary Libben and Gonia Jarema. 2006. *The representation and processing of compound words*. Oxford University Press.

Carlos Ramisch, Helena de Medeiros Caseli, Aline Villavicencio, André Machado, and Maria José Finatto. 2010. A hybrid approach for multiword expression identification. In *International Conference on Computational Processing of the Portuguese Language* , pages 65–74.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 210–218.

Measure of Phonological Reduction

- Goal to establish continuous measure capturing phonological reduction (e.g. loss of secondary stress, consonant cluster reduction)
- Absolute duration not sufficient; need for relative measure of duration
- Compare each compound's final rime (VC(C)) duration to same rime in monosyllabic non-compounds (e.g. *homew**ork*** compared to ***j**erk, cl**erk**, qu**irk***)

Discussion

- Results provide evidence semantic opacity in compounds has reflexes in phonological form
- Cast doubts on categorical notions of compositionality assumed in theoretical aspects of compound representation

Future Directions

- Collect additional compounds from other corpora like BNC and Boston Radio Corpus
- Broaden the number of compounds rated for compositionality

Acknowledgements

Special thanks to friends and colleagues from Cornell Linguistics and the Cornell Phonetics Lab

Contact Information

- http://conf.ling.cornell.edu/forrestdavis/
- fd252@cornell.edu