STATISTICAL LEARNING AND LANGUAGE (IN SPITE OF ARBITRARINESS)

DAVIDE CREPALDI*[1], SIMONA AMENTA[2], and MARCO MARELLI[3]

*Corresponding Author: davide.crepaldi@sissa.it
[1]Neuroscience Area, SISSA, Trieste, Italy
[2]Department of Psychology, University of Gent, Gent, Belgium
[3]Department of Psychology, University of Milano Bicocca, Milano, Italy

1. Introduction

It has long been known that the relationship between form and meaning is generally arbitrary in human languages, that is, forms have no inherent relationships with their meanings (Hockett, 1963). However, it has also been shown that the cumulative cultural evolution of languages does introduce regularities in form–to–meaning mapping (Kirby, Cornish, & Smith, 2008), and that systematicity in this mapping helps learnability, at least in terms of word categorization (Monaghan, Christiansen, & Fitneva, 2011). One apparent end product of this structure–oriented "invisible hand" is linguistic morphology—families of words emerge whose relationship in form predicts their relationship in meaning (e.g., DEAL and DEALER, HUNT and HUNTER). Here we complement this evolutionary evidence with data from Cognitive Neuroscience, showing that the brain codes for these form–to–meaning regularities in a probabilistic way, and uses this information as we process words, either in isolation or embedded in sentence context.

2. Methods

As a test case, we considered the construct developed by Marelli, Amenta, and Crepaldi (2015), *Orthography–to–Semantics Consistency (OSC)*. This is a frequency–weighted average of meaning similarity between any given stem (e.g., DIAL) and all words that include that stem in their orthography (e.g., DIALECT, DIALLED, DIALS, DIALLING, DIALOG, DIALYSIS). Formally:

$$OSC(t) = \frac{\sum_{j=1}^{k} f_{r_x} \cos(\vec{t}, \vec{r_x})}{\sum_{j=1}^{k} f_{r_x}} \quad (1)$$

where t is a stem, f_{r_x} is the frequency of its k orthographic relatives r_x, and $\vec{t}$ and $\vec{r_x}$ are vectorial representations of meaning as extracted from a distributional semantic model (Marelli et al., 2015).

DOI:10.12775/3991-1.017

Essentially, OSC tracks how strongly form similarity correlates with meaning similarity.

This measure was shown to predict word identification times in a large psycholinguistic database (Marelli et al., 2015; Balota et al., 2007). Here we test *OSC* against a behavioural index of sensitivity to form–meaning regularities, i.e., morphological priming; and against neurophysiological data (*Event–Related Potentials, ERP*) collected during natural sentence reading (Frank, Otten, Galli, & Vigliocco, 2015).

3. Results

OSC turns out to qualify morphological priming, in such a way that higher frequency primes have more impact on target processing, either strengthening priming, if they are indeed related to the target (e.g., corns–CORN); or weakening it, if they are not (e.g., corner-CORN).

Brain electrophysiology (the LAN and N400 components) is also shown to be modulated by OSC during sentence reading, at least in words that aren't easily predictable given the available sentence context.

4. Discussion

From a Cognitive Neuroscience point of view, these results suggest some re–thinking of linguistic morphology. Rather than a discrete set of operations over a finite set of well–defined mental representations, morphology can (should?) be seen more generally as part of a form–to–meaning mapping effort carried out by the brain, on the basis of probabilistic knowledge that is accumulated through linguistic experience.

From an evolutionary perspective, these data show that the brain takes advantage of regularities in form–to–meaning mapping in language, thus establishing a psychological/neuroscience counterpart to the progressive emergence of structure through iterated learning. Of course, we don't know whether this phenomenon is an evolutionary reaction of the brain to the independently–triggered emergence of structure in language; or a core feature of the human cognitive machinery, which independently contributed itself to the emergence of structured form–to–meaning mapping. Interestingly, these data also link language learning/evolution/processing to general–purpose cognitive and brain operations (Ellison, 2013).

Acknowledgements

This work is supported in part by an ERC Starting Grant awarded to Davide Crepaldi (679010) and an FWO grant awarded to Marco Marelli, Marc Brysbaert and Simona Amenta (FWO.OPR.2017.0014.01).

References

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. I., Kessler, B., Loftis, B., Neely, J., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*, 445–459.

Ellison, T. M. (2013). Categorisation as topographic mapping between uncorrelated spaces. In *Algorithmic Probability and Friends. Bayesian Prediction and Artificial Intelligence* (p. 131-141). Springer.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1–11.

Hockett, C. F. (1963). The problem of universals in language. In J. H. Greenberg (Ed.), *Universals of Language* (pp. 1–29). Cambridge, Massachusetts: MIT Press.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences, USA*, *105*, 10681–10686.

Marelli, M., Amenta, S., & Crepaldi, D. (2015). Semantic transparency in free stems: The effect of Orthography–Semantics Consistency on word recognition. *Quarterly Journal of Experimental Psychology*, *68*, 1571–1583.

Monaghan, P., Christiansen, M. H., & Fitneva, S. (2011). Balancing arbitrariness and systematicity in language evolution. In *The Evolution of Language* (p. 465-466). World Scientific.