On the Semantics of Non-words and their Lexical Categories

Giovanni Cassani, Yu-Ying Chuang, Harald Baayen
giovanni.cassani@uantwerpen.be, yu-ying.chuang@uni-tuebingen.de, harald.baayen@uni-tuebingen.de

1. Introduction

Children can use the phonological form of a word to infer its likely lexical category [Farmer & al 2006]. In Fitneva & al. [2009], children were presented with non-words which sounded as either English nouns or verbs and asked to match them to an entity or action referent. It was found that the more a non-word sounded like a verb, the more likely the action referent was picked. **But is it about lexical categories or meaning?**

2. Hypothesis

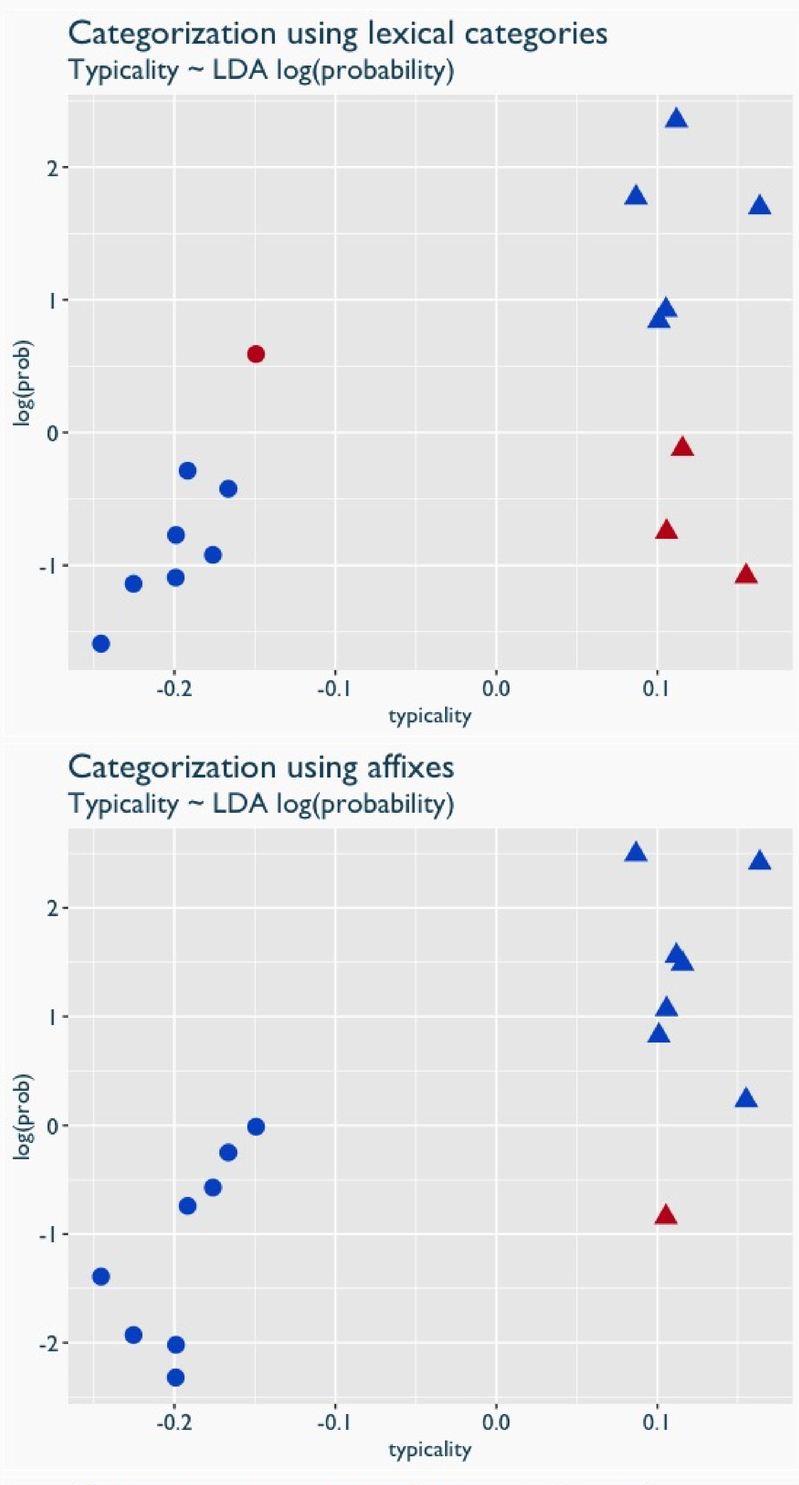
There is a non-arbitrary relation between the sound of a word and its coarse meaning, which children can learn and then use to form semantic impressions given a word form, even out of context.

This semantic impression correlates with semantic knowledge in the lexicon and can be used to cluster new words along semantic distinctions.

3. Methods

- The distributional semantic space of 17826 words was learned using the NDL model on transcribed child-directed speech
- The semantic vectors of the target non-words were generated using the Linear Discrimination learning framework [Baayen & al 2018]
- Non-words were categorized using Linear Discriminant Analysis (LDA) based on:
  - lexical categories of the top 50 semantic neighbors
  - correlations with 43 inflectional and derivational affixes
  - correlations with 20 early-acquired nouns and verbs [Westbury & al 2014]

4. Results

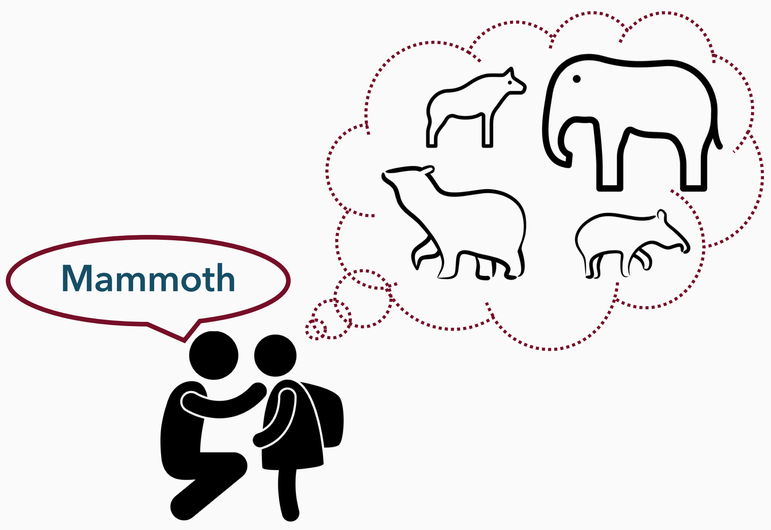


5. Conclusions

It is possible to categorize non-words as either referring to entities or actions purely based on their sounds, even without the mediation of grammar.

Non-words are not meaningless units: they make contact with the lexicon and have coarse semantic content.

The sign isn't entirely arbitrary: in a language, the coarse meaning of a word can be predicted given its sound.



Non-words are just words to be

Acknowledgements

The first author was supported by a BOF/TOP grant (ID 29072) of the Research Council of the University of Antwerp; the second and third authors were supported by a European Research Council grant (#742545 (WIDE)) awarded to the third author. We thank Konstantin Sering and Fabian Tomaschek for the useful discussions.

EBERHARD KARLS UNIVERSITÄT TÜBINGEN
PHILOSOPHISCHE FAKULTÄT
Seminar für Sprachwissenschaft

CLiPS Computational Linguistics & Psycholinguistics
University of Antwerp