NOMEN EST OMEN

Is the word-form of character names a good indicator of their attributes?

Stefano Scola

Department of Cognitive Science and Artificial Intelligence, School of Humanities and Digital Sciences

Introduction

Throughout first language acquisition or when learning a foreign language, we encounter words that we do not know. All of these unknown words are, at least initially, **non-words**. How do we estimate their meaning?

The distributional hypothesis claims that the context in which a word appears is the main source for deriving its meaning. Nevertheless, another plausible explanation is that, when the context is scarce or the vocabulary is poor, **the form of a word** might play an important role in conveying its semantic attributes.

We evaluate different methods in order to extract semantic information from just the form of **characters' names** (with annotation about their gender, age and polarity) to gain insights into **form-based semantic vectors** and to discover whether the way characters are named carries human-like biases (Caliskan, Bryson, & Narayanan, 2017).

Figure 1 illustrates the application of form encoding and form-meaning mapping functions to generate form-based semantic vectors from the form of characters' names

Methods

Word Embeddings:
- word2vec

Word Form Encoding:
- ngrams
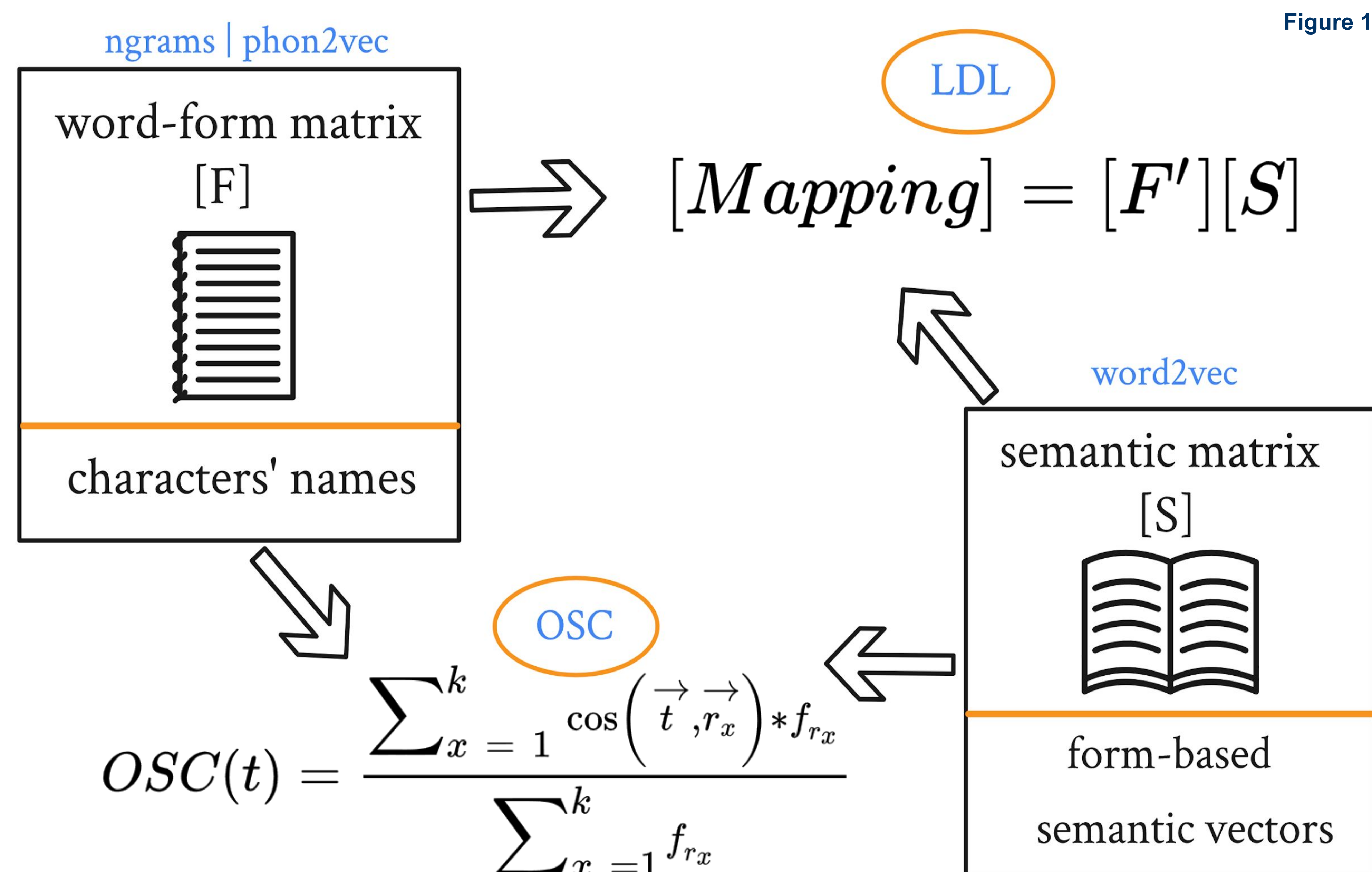- phon2vec

Form-Meaning mapping functions:
- Linear Discriminant Learning (LDL)
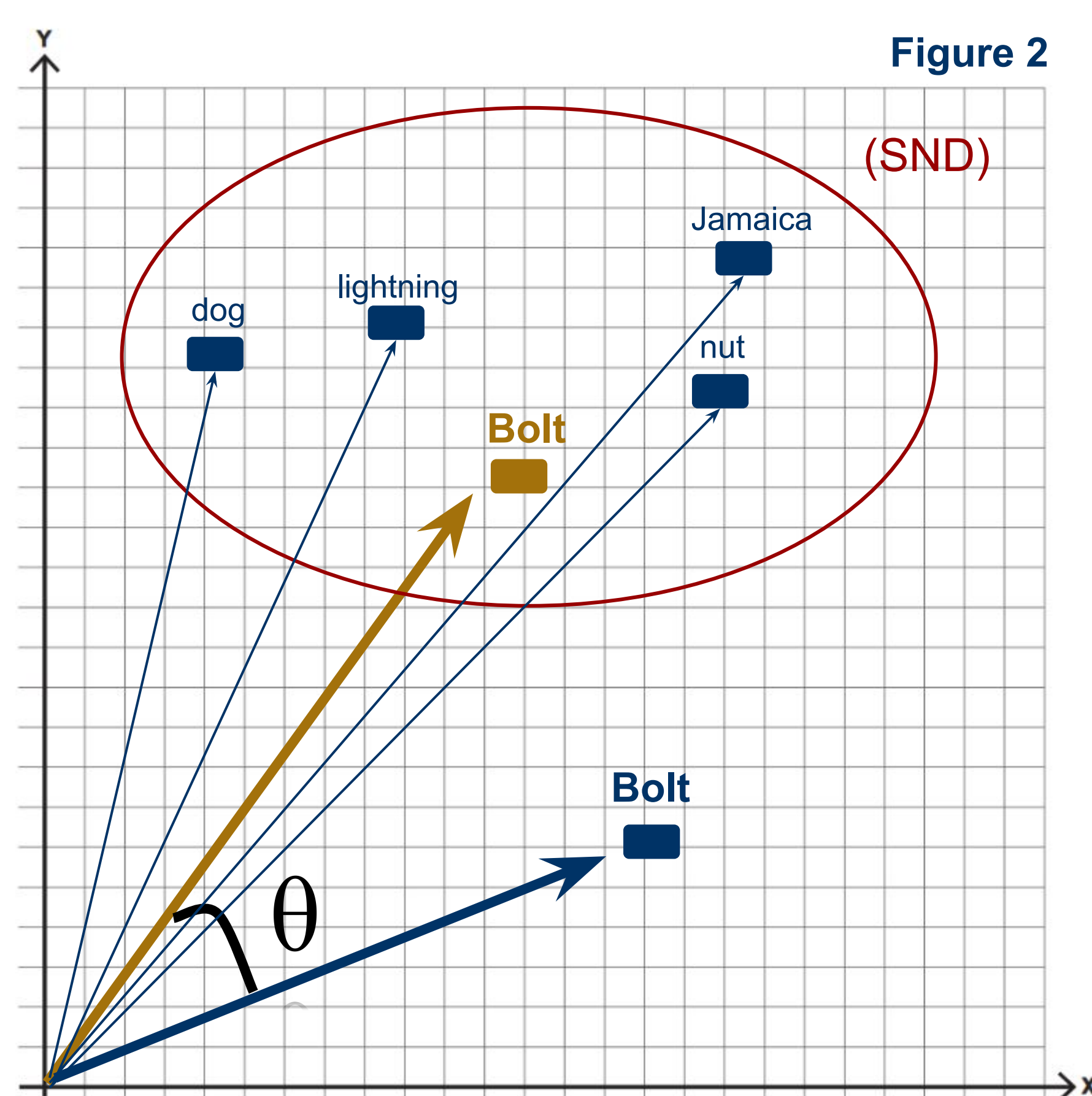- Orthographic Semantic Consistency (OSC)

Corpora
- General Language, Google News
- Fantasy fan fiction stories
- Books for children and young adults

Characters Names
- *Aptronyms* (e.g. Fox, Bolt)
- *Real Names* (e.g. Sam, George)
- *Made-up names* (e.g. Morlok, Araminta)

Figure 1

ngrams | phon2vec

word-form matrix [F]

characters' names

LDL

$$[Mapping] = [F'][S]$$

word2vec

semantic matrix [S]

form-based semantic vectors

OSC

$$OSC(t) = \frac{\sum_{x=1}^{k} \cos\left(\vec{t}, \vec{r_x}\right) * f_{r_x}}{\sum_{x=1}^{k} f_{r_x}}$$

Evaluation

Figure 2

(SND)

Jamaica, dog, lightning, nut, **Bolt**, θ, **Bolt**

Figure 3

F = female; M = male

boy, girl, **family**, **Araminta (F)**, **Morlok (M)**, career

- cosine similarity (*θ*) between **form-based embedding** and **co-occurrence-based embedding** (figure 2).
- *Semantic neighborhood density* **(SND)** of form-based vectors (figure 2).
- cosine similarity between **form-based vectors** of target names and certain attributes of the characters (e.g. boy/girl) (figure 3).
- cosine similarity between **form-based vectors** the target names and vectors that captures human prejudices (e.g. family/career) (Caliskan et al., 2017) (figure 3).
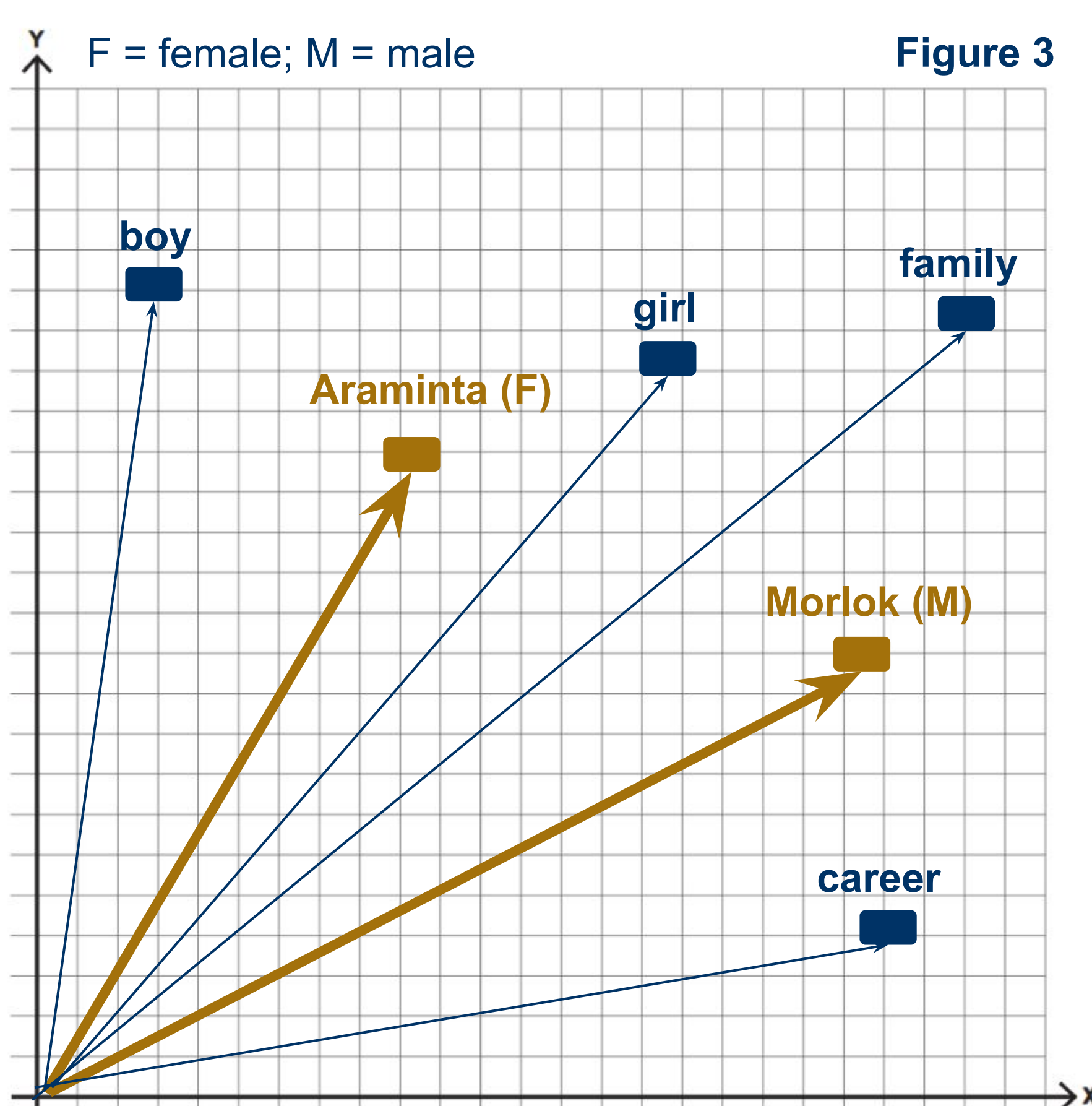
RQs

- To what extent do different **word-form encoding** and **form-meaning mapping** functions capture the sound-symbolism of character names?
- We hypothesise that the interpretation of **made-up** character names is driven by sound-symbolic features (Elsen, 2017), whereas the semantic connotations of **aptronyms** is based on shared language experience. It may be that the sound-symbolic properties of **real names** are inhibited by their episodic content (Sidhu & Pexman, 2019).
- Are character names of early children literature more **sound symbolic** than character names of young adult literature?

References
Elsen, H. (2017). The two meanings of sound symbolism. *Open Linguistics*, *3*(1), 491-499.
Sidhu, D. M., & Pexman, P. M. (2019). The sound symbolism of names. *Current Directions in Psychological Science*, *28*(4), 398-402.
Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183-186.