Uniplex/multiplex pairs and frequency asymmetries in general number languages

Background: This paper examines uniplex/multiplex pairs of nouns across languages. The terms *uniplex* and *multiplex* are used here as notional concepts (Talmy 1988). Uniplex nominals denote a single entity, while multiplex nominals denote a set of multiple entities. In number-marking languages, uniplex and multiplex nominals receive different marking. In English, for example, multiplex nominals are expressed by overt plural forms (e.g., *dog*-ø vs. *dog*-***s***). In Welsh, by contrast, uniplex nominals are overtly marked (singulative/basic forms; e.g., *pys*-***en*** 'pea' vs. *pys*-ø 'peas'). In Haspelmath and Karjus (2017), it was shown that lexemes that crosslinguistically tend to receive singulative marking are more frequently used with a multiplex meaning than with a uniplex meaning, based on the corpus data from five number-marking languages: English, Estonian, Latvian, Norwegian, and Russian.

Research questions: This paper is a replication and extension study of Haspelmath and Karjus (2017) from a contrastive linguistic perspective. We test their hypothesis not only against languages with obligatory number marking but also against general number languages, "in which the meaning of the noun can be expressed without reference to number" (Corbett 2000: 10).

[A] Is Haspelmath and Karjus' (2017) hypothesis replicated in other languages with obligatory number marking, especially in singulative languages such as Sinhala?

[B] Are multiplex-prominent lexemes more frequently used with a multiplex meaning than with a uniplex meaning even in general number languages?

Methods: In order to answer the two above-mentioned questions, we examined large corpora from four number-marking languages (Hindi, Sinhala, Spanish, and Swedish) and five general number languages (Japanese, Korean, Tagalog, Turkish, and Quechua). Following Haspelmath and Karjus (2017), we analyzed the frequencies of 18 lexemes in each language: EAR, LEG, LUNG, GLOVE, SHOE, SKI, APPLE, POTATO, STRAWBERRY, BEE, PIGEON, SHEEP, CHILD, BOY, GIRL, EUROPEAN, AMERICAN, SPEAKER OF (THE RESPECTIVE LANGUAGE). These lexemes are multiplex-prominent in the sense that they crosslinguistically tend to receive singulative marking. In addition, we looked at 18 random lexemes in each language, with the hypothesis that random lexemes would not exhibit the same usage patterns as the 18 multiplex-prominent lexemes.

Coding and annotation: [A] For number-marking languages, we counted the number of basic and derived (i.e., plural or singulative) forms of nouns. [B] For general number languages, we took 40 random samples of each noun from the corpus and manually annotated the counts of (semantically) uniplex and multiplex nouns. To capture the difference between the counts, an "asymmetry index" with a range of −1…1 was used, where negative values indicate dominant singular/uniplex usage, and positive values dominant plural/multiplex usage (Haspelmath & Karjus 2017: 1225).

Results: It is clear from Figures 1 and 2 that, crosslinguistically, the 18 multiplex-prominent lexemes strongly tend to be more frequent in the plural than in the singular (Hindi, Spanish, and Swedish), in the basic than in the singulative (Sinhala), and in a multiplex sense than in a uniplex sense (Japanese, Korean, Tagalog, Turkish, and Quechua), compared to randomly sampled nouns.

Discussion: The above results show that the answer to both [A] and [B] is yes. First, Haspelmath and Karjus' (2017) hypothesis is replicated in Hindi, Sinhala, Spanish, and Swedish and proves to be a robust hypothesis. Importantly, this is the first study to demonstrate that multiplex-prominent lexemes are more frequently expressed by basic forms than by singulative (derived) forms in a singulative language like Sinhala, as Haspelmath and Karjus (2017) predict.

Second, this study also shows that general number languages exhibit the same frequency asymmetries as number-marking languages, even though they lack form distinctions between uniplex and multiplex forms. Thus, by contrasting two different types of languages, this study suggests that frequency asymmetries between uniplex and multiplex nouns universally exist, although they do not always result in coding asymmetries.

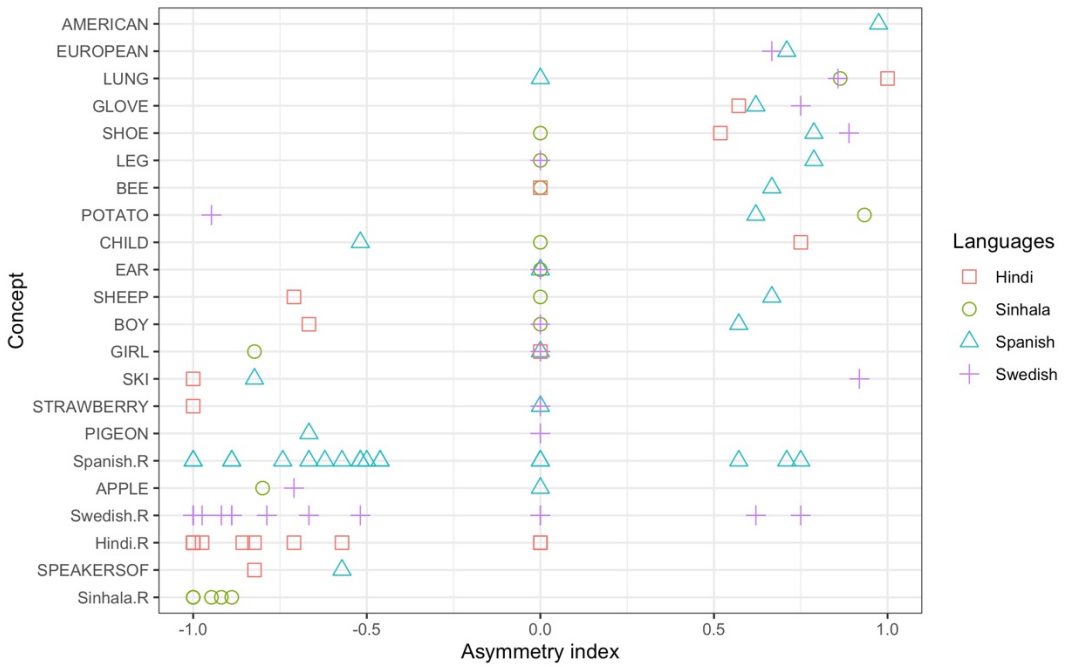

Figure 1: The asymmetry index in number-marking languages ('R' marks the random lexemes)

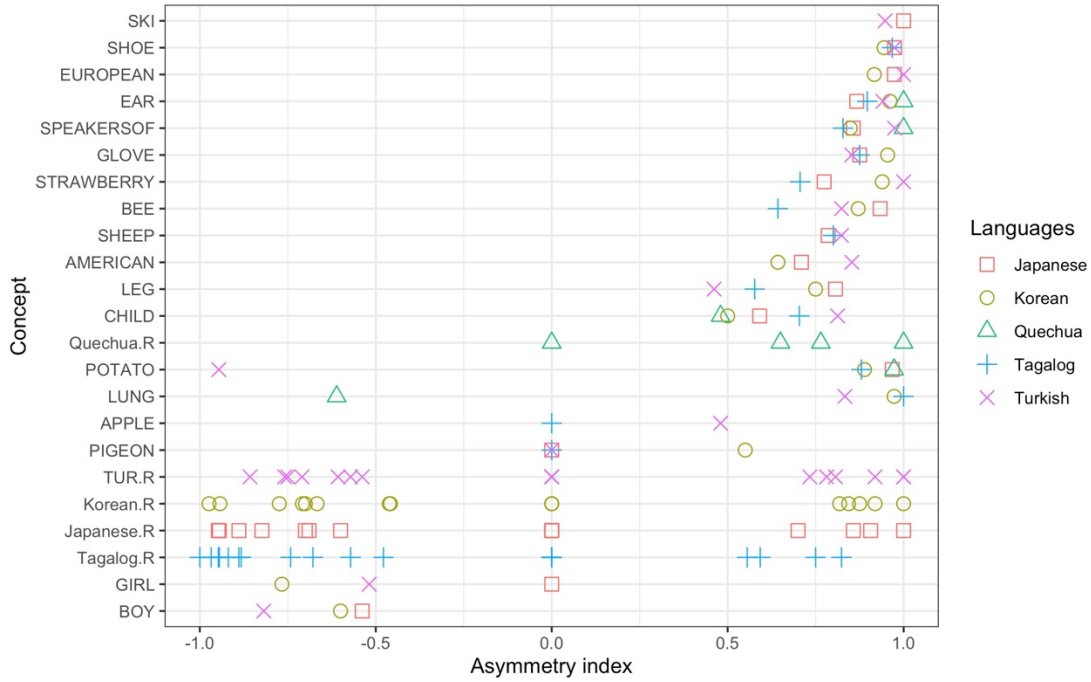

Figure 2: The asymmetry index in general number languages ('R' marks the random lexemes)

References:

Corbett, Greville G. 2000. *Number* (Cambridge Textbooks in Linguistics). Cambridge: Cambridge University Press.

Haspelmath, Martin & Andres Karjus. 2017. Explaining asymmetries in number marking: Singulatives, pluratives, and usage frequency. *Linguistics* 55(6). 1213–1235. https://doi.org/10.1515/ling-2017-0026.

Talmy, Leonard. 1988. The relation of grammar to cognition: A synopsis. In Brygida Rudzka-Ostyn (ed.), *Topics in cognitive linguistics*, 166–205. Amsterdam: John Benjamins. https://doi.org/10.1075/cilt.50.08tal.