Do Noun Classes Have a Semantic Basis? A Multilingual Analysis with Machine Learning

Thomas Esteves Varvella Vicente[1], Ethan Amato[1], Joshua K. Hartshorne[1]

(1) Boston College Department of Psychology and Neuroscience

Lexical gender systems, noun classifier systems, and other noun class systems have long puzzled linguists and laypeople alike. While in some cases the groupings are explicable (names of men having masculine lexical gender), the class of many nouns is puzzling at best. Note that while one can often identify a noun's class from its phonology (Kelly, 1990), we are focused here on whether there is any explanation as to why particular ideas get described by nouns with specific classes (and, if applicable, class-appropriate phonology).

While many studies have investigated semantic correlations of noun classes (e.g., Bergen, 1980; Croft, 1994; Steinmetz, 2006), the sheer number of words involved and the difficulties in quantifying semantics presents a significant stumbling block. To address this, we used word embeddings to quantify semantics, asking whether nouns from different classes occupy distinct or overlapping areas of semantics (for a similar approach, see Williams et al., 2019). This allowed us to consider many thousands of nouns at once for each of 19 languages spanning three different language families (Table 1; see also supplementary page on languages). As a result, findings should be more robust and generalizable, though certainly additional languages and language families must be investigated.

Specifically, using FastText word embeddings (Bojanowski et al., 2017), we used spectral clustering to identify clusters within each class's embedding space before projecting the clusters onto a shared space. We then quantified how much the convex hulls of each cluster overlapped with clusters for other classes. We considered several different metrics (Table 1), all of which indicated a surprising amount of non-overlap between the semantic spaces covered by different noun classes. At the same time, in all languages there was some overlap. The amount of overlap varied across languages, with some languages having a less overlap, like Hebrew, while others have more, like Dutch (Table 1).

We discuss these findings in the context of limitations of the method, including the choice of clustering algorithm, imperfections of FastText embeddings as a measure of semantics, and the availability and quality of the corpora and electronic dictionaries that form our dataset. Nonetheless, it seems clear that noun classes – at least in the three language families studied – while not being completely determined by semantics have a more significant semantic component than some have suspected. We discuss implications for theories of language acquisition and diachronic change.

Supplementary Page for Description of Non-English Languages

The noun class systems in the languages studied here are generally categorized as grammatical gender systems. In these systems, nouns are divided into some number (usually 2 or 3) of genders which are reflected in morphology or agreement. For instance in Spanish, *perro* and *perra* refer to male and female dogs, respectively; adjectives and articles agree with the grammatical gender of the noun:

a. ‘El perro negro’
ART.masc dog black.masc-ending
‘The Black Dog’
b. ‘La perra negra’
ART.feminine dog.feminine black.fem-ending
‘The Black (Female) Dog’

Critically, while grammatical gender sometimes reflects biological sex, it also applies to entities that have no biological sex, like bridges and cars or even abstractions like truth or justice.

Of the languages in our study, all have a masculine and feminine grammatical genders. The languages that also have a neuter gender are Bulgarian, Dutch, German, Icelandic, Polish, Russian, Sanskrit, Serbo-Croatian, Slovak, Telugu, and Ukrainian.

Table 1: Comprehensive Language Data Overview

Language	Family	Tokens	Dimensions	Purity Score	Z Score	Cramer's V
Arabic	Afro-Asiatic	12,631	250	0.52	432.11	0.37
Bulgarian	Indo-European	5,016	238	0.40	138.65	0.33
Dutch	Indo-European	43,768	251	0.26	168.58	0.17
French	Indo-European	63,159	247	0.32	240.46	0.14
German	Indo-European	58,482	245	0.31	367.11	0.19
Hebrew	Afro-Asiatic	6,805	246	0.59	14.35	0.48
Hindi	Indo-European	11,228	253	0.50	136.70	0.27
Icelandic	Indo-European	12,702	252	0.31	179.75	0.24
Italian	Indo-European	113,810	251	0.31	70.88	0.25
Maltese	Afro-asiatic	5,799	197	0.43	78.80	0.27
Polish	Indo-European	70,508	245	0.31	383.23	0.25
Portuguese	Indo-European	49,197	232	0.32	112.25	0.16
Russian	Indo-European	29,894	239	0.33	111.83	0.17
Sanskrit	Indo-European	5,348	187	0.44	213.97	0.32
Serbo-Croatian	Indo-European	30,739	255	0.37	102.32	0.23
Slovak	Indo-European	5,057	239	0.39	107.68	0.33
Spanish	Indo-European	80,661	244	0.34	330.89	0.22
Telugu	Dravidian	4,995	248	0.39	71.10	0.36
Ukrainian	Indo-European	4,274	239	0.37	243.61	0.24

Purity Score *measures the consistency of semantic classifications within clusters.* Z Score *quantifies deviations from the average purity score, indicating the uniqueness of each language's semantic categorization.* Cramer's V *assesses the strength of association between semantic features and noun classes; higher values -> stronger association.*

Bergen, J. J. (1980). The semantics of gender contrasts in spanish. Hispania, 63(1), 48-57.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the association for computational linguistics, 5, 135-146.

Croft, W. (1994). Semantic universals in classifier systems. Word, 45(2), 145-171.

Kelly, M. H. (1992). Using sound to solve syntactic problems: the role of phonology in grammatical category assignments. Psychological review, 99(2), 349.

Steinmetz, D. (2006). Gender shifts in Germanic and slavic: Semantic motivation for neuter? Lingua, 116(9), 1418-1440.

Williams, A., Cotterell, R., Wolf-Sonkin, L., Blasi, D., & Wallach, H. (2019). Quantifying the semantic core of gender systems.