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

Thomas Esteves Varvella Vicente[1], Ethan Amato[1], Joshua K. Hartshorne[2]

[1]Department of Psychology, Boston College [2]Department of Communication Sciences and Disorders, MGH Institute of Health Professions

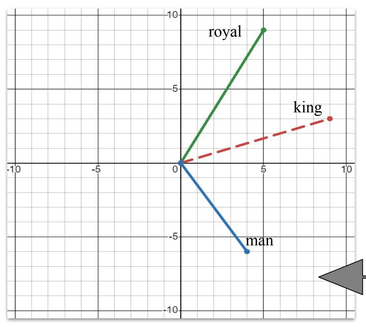
Introduction

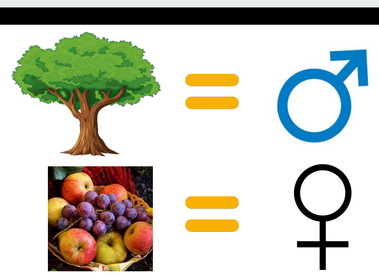
Lexical Gender and Noun Class Systems:

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

Diversity in Noun Class Literature:

- 84% of linguistics literature on English + Indo-European Languages (Kidd & Garcia, 2021)
- Indo-European findings may not generalize to all languages
- Need More Quantitative Studies





Motivating Examples:

- Trees as masculine, fruit as feminine: semantic purity?
- "Masculinidad": Feminine gender defies expectations
- Avoid cross-language analysis due to ambiguity ("bridge")

Word Embeddings:

- High dimension vectors that represent semantic meaning in space (Bojanowski et al. , 2017)
- Encode substring meaning to enhance semantic understanding of different types of languages (e.g. polysynthetic)
- Words closer together share similar meanings
- Allows for vector computation with words

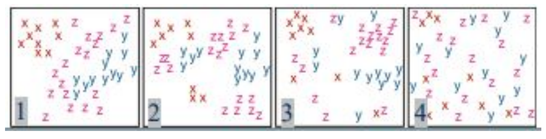
Hypothesis

Extreme vs. Intermediate Hypothesis:

- Extreme hypotheses: No systematicity or complete systematicity
  + Extreme hypotheses don't pass the smell text, but useful for comparison
- Intermediate hypotheses: Classes are systematic but not coherent OR classes have systematic cores plus some random additions.

Hypothesis: Gender in clusters of semantically-close words would follow one of four patterns:

1. Semantically pure
2. Pure but outliers assigned systematically
3. Pure but outliers are assigned randomly
4. No different than chance



Method

(1) Data:

614,073 + gendered nouns being analyzed, from 19 different languages, and 3 different language families.

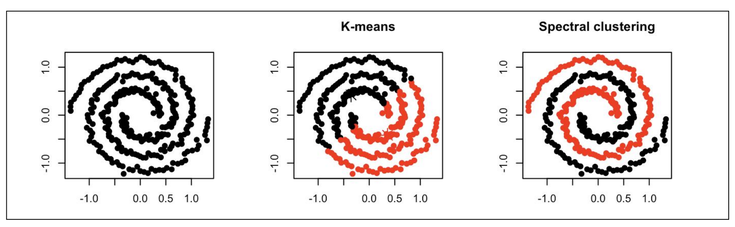
Language	Language Family	Number of Tokens	Vector Dimensions
Arabic	Afro-Asiatic	12,631	250
Bulgarian	Indo-European	5,016	238
Dutch	Indo-European	43,768	251
French	Indo-European	63,159	247
German	Indo-European	58,482	245
Hebrew	Afro-Asiatic	6,805	246
Hindi	Indo-European	11,228	253
Icelandic	Indo-European	12,702	252
Italian	Indo-European	113,810	251
Maltese	Afro-asiatic	5,799	197
Polish	Indo-European	70,508	245
Portuguese	Indo-European	49,197	232
Russian	Indo-European	29,894	239
Sanskrit	Indo-European	5,348	187
Serbo-Croation	Indo-European	30,739	255
Slovak	Indo-European	5,057	239
Spanish	Indo-European	80,661	244
Telugu	Dravidian	4,995	248
Ukrainian	Indo-European	4,274	239

Number of Tokens is the number of words analyzed in the study and vector dimensions is the size of the vectors in the embedding space.

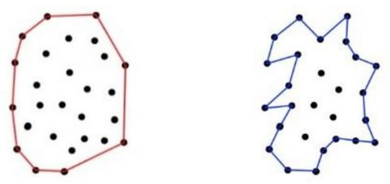
(2) Models:

(Grave et al., 2018)

(3) Spectral Clustering



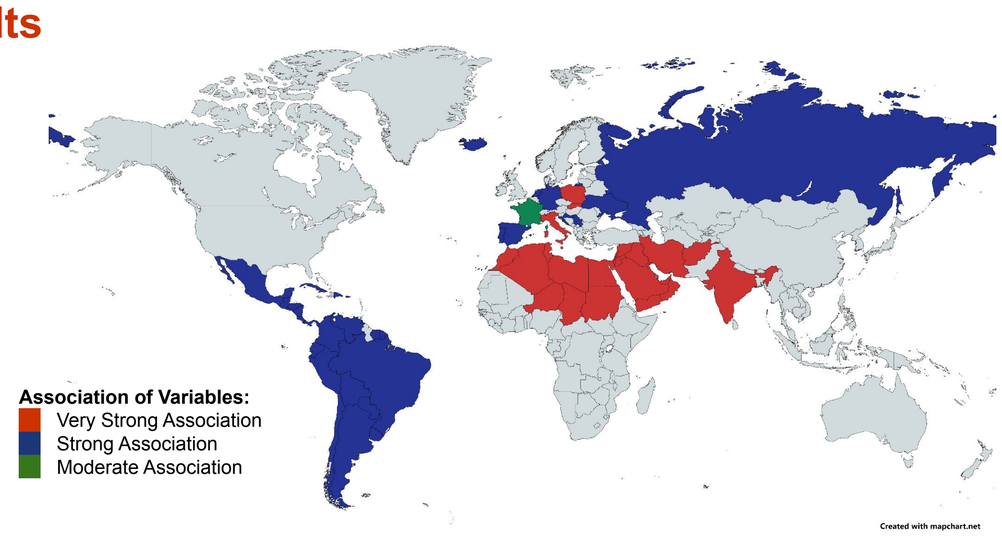
(4) Convex Hull



(a) Convex hull (b) Concave hull

- Counteracts concavity from clustering
- Realistic encapsulation of cluster semantics
- Clustered independently by noun class, then projected to shared space
- Uses the natural shape of the clusters
- Does not force the existence of unwanted clusters

Results



Language	Purity Score	Z Score	Cramer's V
Arabic	0.52	432.11	0.37
Bulgarian	0.40	138.65	0.33
Dutch	0.26	168.58	0.17
French	0.32	240.46	0.14
German	0.31	367.11	0.19
Hebrew	0.59	14.35	0.48
Hindi	0.50	136.70	0.27
Icelandic	0.31	179.75	0.24
Italian	0.31	70.88	0.25
Maltese	0.43	78.80	0.27
Polish	0.31	383.23	0.25
Portuguese	0.32	112.25	0.16
Russian	0.33	111.83	0.17
Sanskrit	0.44	213.97	0.32
Serbo-Croation	0.37	102.32	0.23
Slovak	0.39	107.68	0.33
Spanish	0.34	330.89	0.22
Telugu	0.39	71.10	0.36
Ukrainian	0.37	243.61	0.24

Purity Score: Measures the consistency of semantic classifications within clusters.

Z Score: Indicates the uniqueness of each language's semantic categorization.

Cramer's V: Measures association between semantic features and noun classes; higher values indicate a stronger association.

P Value statistical significance testing underway previous work on smaller datasets strongly outperform chance

Cramer's V > 0.25	Very Strong (hyp 1 & 2)
0.15 < Cramer's V < 0.25	Strong (hyp 2 & 3)
0.10 < Cramer's V < 0.15	Moderate (hyp 3)
Cramer's V < 0.10	Weak (hyp 4)

Conclusions and Future Directions

- Noun class assignments are not semantically pure but are unlikely randomly assigned, suggests and intermediate hypothesis.
- Even with conservative cluster grouping and purity testing, noun class association remains above the weak threshold.
- Most languages show strong or very strong noun class association.
- Study could benefit from a broader range of languages and data.
- Systematicity within languages does not guarantee cross-linguistic systematicity.
- Future testing includes plotting multiple languages in the same embedding space.
- Analyze how cliques of nouns with shared noun class labels track across languages.
- Test the role of animate nouns in systematic categorization.

References

Bergen, J. J. (1980). The semantics of gender contrasts in spanish. Hispania, 63(1), 48-57. doi:10.2307/340811

Bojanowski, Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics, 5, 135–146. https://doi.org/10.1162/tacl_a_00051

Grave, Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning Word Vectors for 157 Languages.

Harris, Z. S. (1954). Distributional Structure. WORD, 10(2–3), 146– 162. doi:10.1080/00437956.1954.11659520

Kidd, E., & Garcia, R. (2021, September 6). How diverse is child language acquisition?. https://doi.org/10.31234/osf.io/jpeyq

Acknowledgements

Thank you to Joshua Hartshorne, Ethan Amato, The Boston College department of Psychology and Neuroscience, and AMLaP for their endless support.