Is Twain's 'Awful German Language' really that Awful? Quantifying the Randomness of Linguistic Gender in 19 Languages

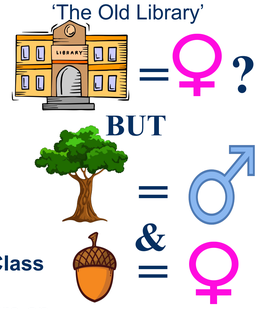
Ethan Amato[1]
Joshua K. Hartshorne[1]
[1]Department of Psychology, Boston College

Background and Motivation

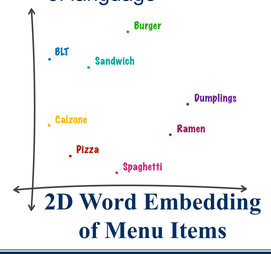
- **Grammatical Gender:**
  - Linguistic systems where nouns assigned 2 gender categories (e.g masc, fem, neut)
- **Unclear Semantics in gender assignment:**
  - Rarely intuitive
  - But also sometimes clearly organized
  - What's going on?

Examples from Spanish:

'La biblioteca vieja
ART.fem library old.fem
'The Old Library'



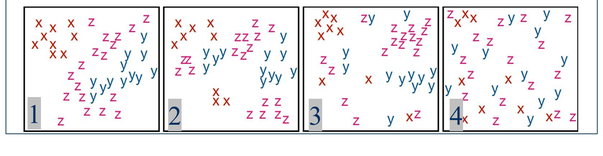
- **Systematic Issues in Noun Class Literature:**
  - 84% of linguistics literature on English + Indo-European languages (Kidd & Garcia, 2021)
  - Gender findings from Indo-Euro Studies may not generalize.
  - Need more quantitative studies of language



- **Word Vectors**
  - Large arrays of numbers that encode meaning. (Bojanowski at al., 2017).
  - Trained w/ *lots* of written text
  - Words closer together in space closer in meaning.

Aims and Hypotheses

- Aim: Quantify relationship between Gender Assignment + Semantics in as many Language Families as possible
- H1: Effect Size between Gender and the Gender of semantically-close words is different than 0.
- H2-5: Gender in Clusters of semantically-close words would follow 1 of 4 patterns:
  1. Semantically Pure
  2. Pure but outliers assigned systematically
  3. Pure but outliers are assigned randomly
  4. No different than chance

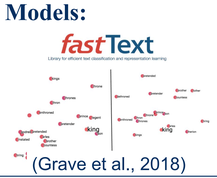


Method

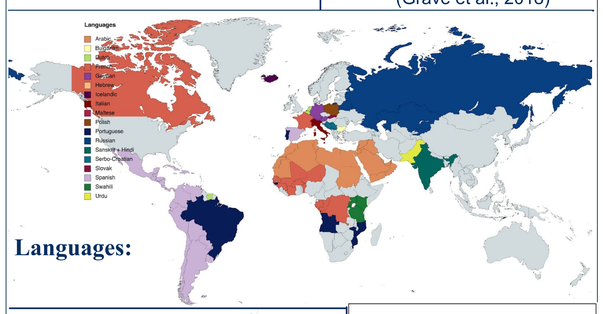
Data:

383,021 Gendered Nouns Analyzed

Models:



(Grave et al., 2018)



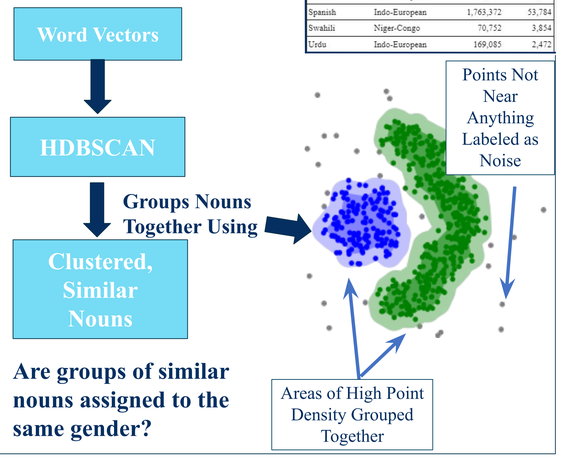
Languages:

K Nearest Neighbor Search:

Is N1's Gender a Good Predictor of N2's?
Do Nouns Usually Share Gender with Closest Neighbor?

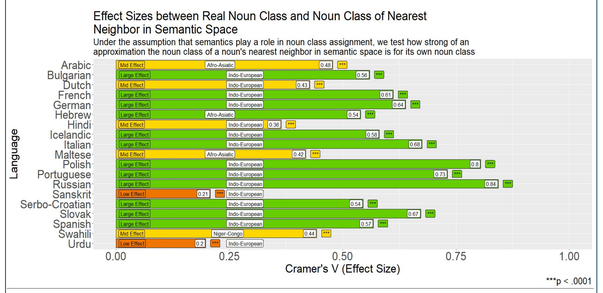
Language	Language Family	# of Wikipedia Articles	N Tokens
Arabic	Afro-Asiatic	1,162,234	11,538
Bulgarian	Indo-European	280,446	4,749
Dutch	Indo-European	208,602	28,862
French	Indo-European	2,410,118	43,686
German	Indo-European	2,676,801	42,829
Hebrew	Afro-Asiatic	313,333	5,914
Hindi	Indo-European	153,538	9,238
Icelandic	Indo-European	54,121	11,019
Italian	Indo-European	1,747,435	66,038
Maltese	Afro-Asiatic	4,529	3,803
Polish	Indo-European	1,516,037	25,898
Portuguese	Indo-European	1,088,035	5,318
Russian	Indo-European	180,6335	26,488
Sanskrit	Indo-European	11,644	3,490
Serbo-Croatian	Indo-European	456,423	30,218
Slovak	Indo-European	239,950	3,823
Spanish	Indo-European	1,763,372	53,784
Swahili	Niger-Congo	70,752	3,854
Urdu	Indo-European	169,085	2,472

HDBSCAN Clustering:



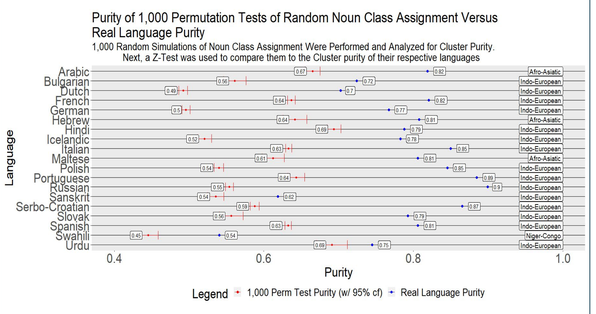
Results

Nearest Neighbor Search

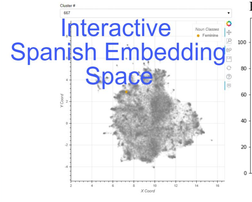


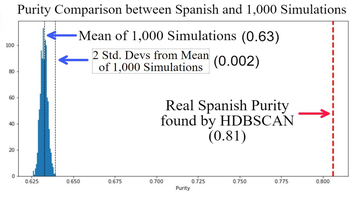
- 17/19 Languages have mid-high effect sizes, but all languages had some degree of systematicity

HDBSCAN:



- All Languages ***STRONGLY*** outperform 1,000 simulations of random assignment





Conclusions

- Noun Class assignment not completely pure nor completely arbitrary.
- Amount of noise detected by HDBSCAN points towards random assignment of outliers but further testing necessary
- Systematicity between Semantics and Noun class across multiple language families implies cognitive predisposition
- Future testing will include plotting multiple languages onto a common embedding space to see if different languages categorize the same concepts consistently

All Interactive Visualizations uploaded to:
https://ardianor-ops.github.io/semanticorewebsite/

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 Yujing Huang, Joshua Hartshorne, Wesley Ricketts, Seun Eisape, Aidas Aglinskey, Damian Blasi, and Katie McCauliffe for their endless support and assistance with this project.