Using similarity measures to capture form-function diversities in P-demotion domain

Mohammad Tavakoli
Adam Mickiewicz University in Poznan

Keywords: P-demotion, cosine similarity, similarity measures, cluster analysis, statistical analysis

P-demotion domain is the main focus of this study which involves a family of valency and voice alternations in which the P-argument of a transitive construction loses its properties as a core argument without affecting the argument structure of the construction, hence agent remains agent and patient remains patient. (Janic and Witzlack-Makarevich 2021; Zúñiga and Kittilä 2019). P-demotion as an operation triggers intransitive constructions in which P is expressed as an oblique, incorporated, suppressed, or omitted. Such constructions are discussed under various terms in the literature including antipassive (1), conative, noun incorporation, and A-lability.

1) Warrungu (Pama-Nyungan; Tsunoda 1988: 598)

a)	*pama-ngku*	*kamu-Ø*	*yangka-n*
	man-ERG	water-ABS	search-PP

b)	*pama-Ø*	*kamu-wu*	*yangka-kali-n*
	man-ABS	water-DAT	search-ANTIP-PP

both:‘A man looked/looks for water.’

Here, my main goal is to capture the diversity of P-demotion clauses in the world’s languages based on their formal and functional characteristics using a statistical approach. To do so, we coded 55 genealogically different languages from six macro areas, represented as a matrix, where rows stand for languages and columns for both formal features (including indexation, flagging, voice marking) and functional ones (referentiality, affectedness, etc.). The rows of this matrix, here the languages of the database, can be represented as vectors. By turning them into vectors I will be able to compare them via various existing similarity measures, e.g. cosine similarity (Glynn and Robinson 2014; Levshina 2015).

By applying a similarity measure, I will produce a distance matrix in which the values show dis(similarities) between the languages of the database. Since there are 55 languages, the output will be a high-dimensional 56×56 matrix. To further interpret the results of the distance matrix, we use cluster analysis (Glynn and Robinson 2014; Gries 2009; Husson, Le, and Pagès 2010); that is categorizing languages of the database based on how close they are in the distance matrix. we expect languages with the same P demotion realization to cluster together. The analysis will help us determine the correlations between form and function in P-demotion domain since formal and functional features are used to build the vectors in the first place. Clustering languages in this way will further allow us to investigate the scope and diversity of P demotion realizations across languages of the database and to explore if genealogical and geographical affinity are influential factors in the clustering of similar languages.

References

Glynn, Dylan, and Justyna Robinson. 2014. *Corpus Methods for Semantics*. *Hcp.43*. Amsterdam/Philadelphia: John Benjamins. https://benjamins.com/catalog/hcp.43.

Gries, Stefan Th. 2009. “Statistics for Linguistics with R: A Practical Introduction.” In *Statistics for Linguistics with R*. Berlin/Boston: De Gruyter Mouton.

Husson, Francois, Sebastien Le, and Jérôme Pagès. 2010. *Exploratory Multivariate Analysis by Example Using R*. New York: CRC Press.

Janic, Katarzyna, and Alena Witzlack-Makarevich. 2021. *Antipassive: Typology, Diachrony, and Related Constructions*. Amsterdam: John Benjamins.

Levshina, Natalia. How to Do Linguistics with R: Data Exploration and Statistical Analysis. Amsterdam/Philadelphia: John Benjamins, 2015.

Tsunoda, Tasaku. 1988. “Antipassives in Warrungu and Other Australian Languages.” In *Passive and Voice*, edited by Masayoshi Shibatani, 595–650. Amsterdam: John Benjamins.

Zúñiga, Fernando, and Seppo Kittilä. 2019. *Grammatical Voice*. Cambridge: Cambridge University Press.