An automated classification of idiomatic wh-pseudoclefts

Hollen Foster-Grahler– Washington State University

English wh-pseudoclefts (WHPCs) are a type of clefted construction headed by a wh-word. They are defined primarily by their information structure: wh-word + *topic* + copula + *focus* (Prince, 1978). The topic consists of given information, often from earlier in the discourse, while the focus contains information that the speaker believes is new to the listener. WHPCs are most common in spoken language; as such, there is significant variation in their production. This is especially apparent in the realization of the copula, which is frequently omitted due to interruption or ellipsis (Collins, 1994).

Previous studies have attempted to define the WHPC abstractly in terms of syntax, semantics, or discourse function (Yoo, 2003; Hopper, 2001), or describe their use in corpora (Weinert & Miller, 1994; Zhou, 2021). However, it is rarely acknowledged that the actual structure of the WHPC is that of a lexically opened idiom. As per Fillmore, lexically opened idioms fulfill pragmatic and semantic roles beyond the sum of their parts and have features of a preset structure, but are not an unalterable string of words (1988). As such, when analyzing the structure or use of WHPCs, idiomaticity should be forefront. This begs the questions: how do WHPCs display idiomaticity? How can we analyze the degree of variation within a WHPC? And can that variation predict the idiomatic type?

In the project at hand, I used regular expressions to traverse the British National Corpus's spoken language subcorpus to find data on how people use WHPCs in everyday speech. From this data, I developed a typology based on Prince's (1978) that focuses on discursive function– the pragmatic use of each type of idiomatic WHPC. I found these functional categories shared surface-level characteristics like syntax, verb choice, and complexity. To encode these superficial features, I created an eight-dimensional vector, the dimensions of which range from 0 (most prototypical/common) to 1 (most infrequent). I am in the process of using a single value decomposition (SVD) to project the vectors into fewer dimensions and see if it's possible to cluster them into the seven discursive categories. If the SVD fails to yield significant results, I will use the kernel trick with a support vector machine to see if a non-linear division better classifies the vectors. I will examine the accuracy using the leave-one-out cross validation.

If either of these methods reasonably classifies the WHPCs, it would provide the framework for a construction of wh-pseudoclefts (construction in the Construction Grammar sense of "a symbolic mapping between form and meaning" (Dunn, 2023)). If it does not, that will speak to the non-atomic emergence of meaning in idiomatic expressions, indicating that quantitative analysis requires a more delicate hand at coreferentiality to identify the idiomatic structure (Westerstahl, 2002). Either way, we will have another step forward in understanding the interface between form and function in lexically opened idiomatic constructions.

References:

Collins, P. (1994). *Cleft and Pseudocleft Constructions in English*. Routledge, London.

Dunn, J. (2023). Syntactic variation across the grammar: modelling a complex adaptive system. *Frontiers in Complex Systems*. https://doi.org/10.3389/fcpxs.2023.1273741

Fillmore, C. J., Kay, P., & O'Connor, M. C. (1988). Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, *64*(3), 501–538. https://doi.org/10.2307/414531

Hopper, P. J. (2001). Grammatical constructions and their discourse origins: prototype or family resemblance? *Applied Cognitive Linguistics I: Theory and Language Acquisition, Putz, Martin, Niemeier, Susanne, & Dirven, Rene (eds.), Berlin: Mouton de Gruyter* 19(1), pp. 109-129 (Originally published 2001, pp. 109–130). De Gruyter, Inc. https://doi.org/10.1515/9783110866247.109

Prince, E. (1978). A Comparison of Wh-Clefts and it-Clefts in Discourse. *Language*, 54(4), pp. 883-906.

Weinert, R. & Miller, J. (1996). Cleft constructions in spoken language. *Journal of Pragmatics*, 25(2), pp. 173-206. https://doi.org/10.1016/0378-2166(94)00079-4

Westerstahl, D. (2002). On the compositionality of idioms: An abstract approach. In D. Barker-Plummer, D. Beaver, J. van Benthem, P. Scotto di Luzio (Eds.), *Proceedings of LLCS,* CSLI Publications.

Yoo, E.J. (2003). Specificational Pseudoclefts in English. In Stefan Müller (Ed.), *Proceedings of the HPSG03 Conference*, CSLI Publications.

Zhou, H. & Chen, M. (2021). What Still Needs to be Noted: Pseudo-Clefts in the Academic Discourse of Applied Linguistics. *Frontiers in Psychology*, Vol. 12. https://doi.org/10.3389/fpsyg.2021.672349