

Distributional investigation into the (in)homogeneity of grammatical categories: an Agreement Groups analysis

László Drienkó
SzSzC Jáky, Székesfehérvár, Hungary
drienko.laszlo@jaky.hu

In this study we assume that grammatical categories in a language are, to a great extent, determined by distributional properties of words, and, in particular, that linguistic constructions can define grammatical categories (cf. e.g. Croft, 2023). Specifically, we analyse the *agreement categories* of the *Agreement Groups* (AG) framework as described in e.g. Drienkó (2014, 2020) since AGs can be viewed as representing cognitive processing units for linguistic constructions or cognitive building blocks for such constructions.

Distributional methods, as pioneered e.g. in Harris (1951, 1952) exploit intrinsic statistical information to gain insight into the structure of language. *Context* representing the combinatorial properties of individual words is a critical notion in determining their overall distribution. Some authors emphasise the role of specific contexts, *frames*, in word categorisation (e.g. Mintz 2003, Weisleder and Waxman 2010, St. Clair et al. 2010, Wang and Mintz 2010). AGs can be viewed as combinations or superimpositions of such framing contexts.

AGs are groups of utterances differing from a base utterance in exactly one word and allow for combinatorial generalisation to novel utterances. The utterance positions within AGs represent agreement categories (more precisely: *AG categories*) which may correspond with grammatical or semantic categories. However, these AG categories are partial or incomplete in the sense that they constitute subsets of overall linguistic categories such as Noun, Verb, etc. and they may contain words from multiple categories. The present work quantifies how homogeneous, or in other words, how mixed AG categories are.

For measuring homogeneity within an AG category we used the maximal number of those words within the category that were assigned the same grammatical category by a part-of-speech (POS) tagger, divided by the number of all the words in the category. Average homogeneity was calculated as the sum of the homogeneity values for all AG categories divided by the number of all AG categories.

The collection of AGs came from the mother-child data of the Anne sessions in the Manchester corpus (Theakston et al. 2001) in the CHILDES database (MacWhinney 2000). There were 60516 AG categories altogether, of which only categories containing at least two words were considered. Thus, the number of AG categories analysed was 17136. The maximal number of words in a category was 137. The POS tagger assigned 30 POS categories to the 2506 word types in the AGs. We found that the average homogeneity of AG categories was 55%.

In an additional analysis we checked how utterance length affects the categorisation capacity of AGs, viz. the homogeneity of AG categories. Since an AG contains utterances of either 2 or 3 or 4 or 5 words, we calculated homogeneity for the respective cases. Again, only categories containing at least two words were considered. The results are summarised as Table 1.

Utterance length in AG (words)	2	3	4	5
Average homogeneity of categories (%)	41	59	69	75

Table 1.

Our analysis is in line with a view that word classes are construction-specific and demonstrates how a distributional model, specifically the AG framework, allows for direct quantification of the homogeneity of grammatical categories. The 55% average value shows that, overall, the AG categories obtained from the analysed mother-child language contain a considerable degree of inhomogeneity. However, longer utterances seem to be more concentrated with respect to POS categorisation, as AG categories in longer utterances are more homogenous. We note that AG categories also licence a POS vector-space interpretation: for a given AG category we can obtain its vector components by counting how many times a particular POS occurs in the AG category (cf. e.g. Stolcke, 1991).

REFERENCES

- Croft, W. (2023). Word Classes in Radical Construction Grammar. In van Lier, E. (ed.) *The Oxford Handbook of Word Classes*, Oxford Handbooks
<https://doi.org/10.1093/oxfordhb/9780198852889.013.48>
- Drienkó, L. (2014). Agreement groups analysis of mother-child discourse. In Rundblad, G., Tytus, A., Knapton, O., and Tang, C. (eds.) *Selected Papers from the 4th UK Cognitive Linguistics Conference*. London: UK Cognitive Linguistics Association. pp. 52-67.
- Drienkó, L. (2020). Agreement Groups and dualistic syntactic processing. In Haselow, A. and Kaltenböck, G. (eds.) *Grammar and Cognition: Dualistic models of language structure and language processing, [HCP 70]*. 310-354. John Benjamins Publishing Company.
- Harris, Z. S. (1951). *Methods in structural linguistics*. Chicago, IL, US: University of Chicago Press.
- Harris, Z. S. (1952). Discourse analysis. *Language*, 28(1). 1-30.
- MacWhinney, B. 2000. *The CHILDES Project: Tools for analyzing talk*. 3rd Edition. Vol. 2: The Database. Mahwah, NJ: Lawrence Erlbaum Associates.
- Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, Volume 90, Issue 1, 91-117. doi:10.1016/S0010-02770300140-9 |
- St. Clair, M. C., Monaghan, P., and Christiansen, M. H. (2010). Learning grammatical categories from distributional cues: Flexible frames for language acquisition. *Cognition*. Volume 116, Issue 3, 341-360.
- Theakston, A. L., Lieven, E. V., Pine, J. M., and Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: an alternative account. *Journal of Child Language*. 28(1):127-52.
- Wang, H., Mintz, T. H. (2010). From Linear Sequences to Abstract Structures: Distributional Information in Infant-direct Speech. In Jane Chandlee, Katie Franich, Kate Iserman, Lauren Keil (eds.) *Proceedings Supplement of the 34th Boston University Conference on Language Development*
- Weisleder, A, and Waxman, S. R. (2010). What's in the input? Frequent frames in child-directed speech offer distributional cues to grammatical categories in Spanish and English. *Journal of Child Language*. Nov; 37(5):1089-108. Epub. 2009 Aug 24.

Stolcke, A. (1991). Syntactic Category Formation with Vector Space Grammars. *Proceedings of the 13th Annual Conference of the Cognitive Science Society*, University of Chicago, Chicago, IL, 908–912.