Predicting compound branching directions with distributional semantics

Annika Schebesta[1], 1, Jessica Nieder[2], Motoki Saito [3]

[1] Universität Siegen; [2]University of Malta,[3] Carl von Ossietzky Universität Oldenburg

annika.schebesta@uni-siegen.de

The default branching direction in English triconstituent nominal compounds (NNN) is reported to be left-branching [1, 2, 3]: left-branching NNN of the structure [N1N2] N3 ([*seat*$_{N1}$ *belt*$_{N2}$] *law*$_{N3}$) are more frequent than right-branching NNN of the structure N1 [N2N3] (*corner*$_{N1}$ [*drug*$_{N2}$*store*$_{N3}$]). Generally, the branching of NNN is simultaneously provided by several determinants: the order of composition, the lexical bigram frequency of N_1N_2 and N_2N_3, the spelling of the NNN, and the meaning that is conveyed with the NNN have variant power of influencing the parsing, and frequently are in competition about branching direction. The meaning of semantically transparent compounds like *coffee bean roaster* can be inferred from its constituents, while the meaning of more opaque compounds like *horseradish dip* is less predictable. Thus, parsing the branching is a complex cognitive process in the human mind, informed by multiple sources of linguistic and contextual information. In this study, we investigate whether it is possible to leverage distributional semantics to predict branching computationally.

We used two sets of data: one data set with 465 NNN from BURSC [4] and one data set with 100 NNN originally constructed for a production experiment [3]. The branching direction of the corpus data was determined in a semantic and orthographic analysis by two raters [5]. The experimental data contains 50 semantically ambiguous NNN that are set in two different contexts each which suggest one branching direction over the other, resulting in 50 left-branching and 50 right-branching NNN. The resulting 100 NNN have been rated for their branching direction by 46 native speakers in an online experiment.

We retrieved contextual embeddings using the pre-trained uncased BERT base model (`bert-base-uncased`). For each NNN, we extracted the full context sentence in which the compound appeared (from corpus text files or from the experimental stimuli). For each noun within the compound (N_1, N_2, N_3), we located its occurrence in the unmasked sentence and obtained its corresponding token-level embedding from the last hidden layer of BERT. When a noun was represented by multiple subword tokens, we averaged the respective token vectors. This procedure yields three contextualized embeddings per NNN: one for the free noun and one for each embedded constituent, representing each noun's distributional semantics in its actual sentence context.

These embeddings were then used as input features in a Linear Discriminant Analysis (LDA) to assess whether branching direction could be predicted based on distributional semantic patterns. Table 1 summarizes classification performance of the LDA model. Overall, the LDA classifier achieved high accuracy on the training data, however, when presented with held-out data we observe a substantial drop of accuracies (Corpus: 0.68%, Experiment: 0.23%). Despite the low accuracy on held-out data, left-branching compounds were classified more accurately during training, reflecting both the overall left-branching preference in English NNN compounds and participants' behavioural preferences. Interestingly, this pattern reversed in the experimental test data: while the in-bag results showed a bias toward left-branching compounds, out-of-bag predictions favored right-branching ones (RB $>$ LB), indicating that the model's apparent left-branching advantage did not generalize.

This highlights the ambiguous nature of the compounds in the experimental data, showing no clear preference by either the participants or the model.

These findings highlight both the potential and the limitations of using distributional semantic models like BERT for modelling human parsing decisions in structurally ambiguous compounds. While contextual embeddings capture some cues relevant for branching, reliably predicting human-like interpretation patterns, especially for less frequent structures, remains a key challenge for computational models of language processing.

	In-the-bag	**Out-of-bag**
Corpus	Acc: 0.983 (Baseline: 0.705)	Acc: 0.677 (10-fold CV)
	Mis-predicted LB→RB: 4; RB→LB: 4	Random 10% held-out (100×): LB = 0.761, RB = 0.528
	Bias towards LB: 0.50 (LB:RB = 50:50)	
Experiment	Acc: 0.740 (Baseline: 0.500)	Acc: 0.230 (10-fold CV)
	Mis-predicted LB→RB: 11; RB→LB: 15	Random 10% held-out (100×): LB = 0.279, RB = 0.307
	Bias towards LB: 0.577 (LB:RB = 15:11)	Random 10% held-out (100×10): LB = 0.270, RB = 0.298
		$t(17.79) = -4.211$, $p = 0.00054$ (RB $>$ LB)

Table 1: Summary of classification accuracy and prediction patterns for Corpus and Experiment datasets (In-the-bag and Out-of-bag) based on *unmasked tokens*.

References

[1] Schebesta, A., & Kunter, G. (2022). Constituent durations in English NNN compounds: A case of strategic speaker behavior? *Journal of Phonetics*, *94*, 101164. https://doi.org/10.1016/j.wocn.2022.101164

[2] Huber, E. (2023). *Tri-Constituent Compounds. A Usage-Based Account of Complex Nominal Compounding*. De Gruyter Mouton. https://doi.org/doi:10.1515/9783111081694

[3] Schebesta, A. (2024). *NNN compounds in English: Investigating the interface of morphology, lexical frequency and the phonetic signal*. [Unpublished PhD dissertation].

[4] Ostendorf, M., Price, P. J., & Shattuck-Hufnagel, S. (1995). The Boston University Radio News Corpus. https://catalog.ldc.upenn.edu/docs/LDC96S36/

[5] Kösling, K., & Plag, I. (2009). Does branching direction determine prominence assignment? An empirical investigation of triconstituent compounds in English. *Corpus Linguistics and Linguistic Theory*, *5*(2), 201–239. https://doi.org/10.1515/CLLT.2009.009