DISTRIBUTION-BASED PREDICTION OF THE DEGREE OF GRAMMATICALIZATION FOR GERMAN PREPOSITIONS

DOMINIK SCHLECHTWEG* and SABINE SCHULTE IM WALDE

*dominik.schlechtweg@gmx.de
Institute for Natural Language Processing, University of Stuttgart

1. Introduction

Grammaticalization refers to the diachronic "development from lexical to grammatical forms, and from grammatical to even more grammatical forms" (Heine & Kuteva, 2007, p. 32). It is assumed to go along with *desemanticization*, a process of losing descriptive meaning (Heine, 2003; Bybee, 2015; Lehmann, 2015; Heine & Kuteva, 2007). For instance, the German noun *Trotz* 'defiance' acquired a less specific, grammatical meaning as preposition, cf. *trotz des Sturms* 'despite the storm', in which the original descriptive meaning is lost. The resulting prepositional meaning is semantically more general and often highly polysemous (Di Meola, 2014, cf. p. 134, 151). In addition, grammatical categories show a high degree of obligatoriness (Di Meola, 2014, cf. p. 39f.), i.e., they must be specified in a sentence (Lehmann, 2015, cf. p. 14). These three highlighted properties of grammaticalized expressions (generality, polysemy, obligatoriness) have a directly observable impact on their contextual distributions: they are used in a greater number of contexts (Weeds & Weir, 2003; Heine & Kuteva, 2007; Santus et al., 2014; Bybee, 2015; Schlechtweg et al., 2017). Together with the assumption that grammaticalization is a continuous process (Di Meola, 2014, cf. p. 68), these observations motivate our central hypothesis:

Hypothesis The degree of grammaticalization of an expression correlates with the unpredictability of its context words (contextual dispersion).

2. Method

In computational linguistics, a prominent corpus-based measure of contextual dispersion is word entropy (Hoffman et al., 2013; Santus et al., 2014). We exploit this measure in order to test our central hypothesis. First, we create a test set of German prepositions with different degrees of grammaticalization; we then (i) compute Spearman's rank-order correlation coefficient (ρ) between test set and word entropy scores, and (ii) use Average Precision (AP) to measure how well the scores distinguish between degrees.

For computing word entropy we induce a Distributional Semantic Model with window size 2 from a part of the SdeWaC corpus (Faaß & Eckart, 2013) with

DOI:10.12775/3991-1.105

approx. 230 million tokens. Low-frequency and functional words are deleted, and every token is replaced by its lemma plus POS-tag. For comparison, we also compute other quantitative measures of different aspects of contextual dispersion: word frequency and the number of context types.[1]

3. Test Set

Di Meola (2014) distinguishes between (i) prepositions with the form of a content word (e.g., *trotz*), (ii) prepositions with the form of a syntactic structure (e.g., *am Rande*) and (iii) prepositions with the form of a function word (e.g., *vor*). Prepositions in (i) and (ii) show a low to medium degree of grammaticalization, while the ones in (iii) show a high degree (cf. p. 60). We focus on prepositions with the form of a PP from (ii), because Di Meola provides fine-grained distinctions, and (iii), to exploit a wide range of degrees. The final test set contains 206 prepositions with four degrees of grammaticalization (1: low – 4: high).[2]

4. Results

Table 1 shows that there is indeed a moderate correlation between entropy and the degree of grammaticalization, but frequency and the number of context types outperform entropy. Frequency has the highest overall correlation with grammaticalization: it is only .01 different to the correlation of context types ($p = .6$, two-tailed, Steiger's Z-test), but with .04 clearly different from entropy ($p = .06$). Frequency also distinguishes best between most of the degree levels; context types are generally comparable and in one case even the best predictor. Overall, the table clearly demonstrates that the more different the degrees of grammaticalization are, the better they are distinguished by the three measures.

Table 1. Results for predicting degrees of grammaticalization.

	entropy	frequency	types
AP (degrees 1 vs. 2)	0.54	**0.56**	0.55
AP (degrees 1 vs. 3)	0.67	**0.68**	**0.68**
AP (degrees 1 vs. 4)	0.89	**0.92**	**0.92**
AP (degrees 2 vs. 3)	0.67	**0.69**	0.68
AP (degrees 2 vs. 4)	0.89	**0.92**	**0.92**
AP (degrees 3 vs. 4)	0.84	0.87	**0.88**
Spearman's ρ (rank)	0.42	**0.46**	0.45

Our findings contribute an empirical perspective to the relationship between grammaticalization and frequency, which has been discussed intensively (e.g., Di Meola, 2014, cf. p. 173) but not been investigated in a rigorous way, as done here.

[1] Code: `https://github.com/Garrafao/MetaphoricChange`.

[2] The test set is provided together with the predicted measure scores as supplementary material.

References

Bybee, J. L. (2015). *Language change.* Cambridge, United Kingdom: Cambridge University Press.

Di Meola, C. (2014). *Die Grammatikalisierung deutscher Präpositionen* (2 ed.). Tübingen: Stauffenburg.

Faaß, G., & Eckart, K. (2013). SdeWaC – A corpus of parsable sentences from the web. In I. Gurevych, C. Biemann, & T. Zesch (Eds.), *Language processing and knowledge in the web* (Vol. 8105, p. 61-68). Springer Berlin Heidelberg.

Heine, B. (2003). Grammaticalization. In *The Handbook of Historical Linguistics* (pp. 575–601). Oxford: Blackwell Publishing Ltd.

Heine, B., & Kuteva, T. (2007). *The genesis of grammar: A reconstruction.* Oxford University Press.

Hoffman, P., Lambon Ralph, M., & Rogers, T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods*, *45*(3), 718–730.

Lehmann, C. (2015). *Thoughts on grammaticalization.* Language Science Press.

Santus, E., Lenci, A., Lu, Q., & Schulte im Walde, S. (2014). Chasing hypernyms in vector spaces with entropy. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 38–42).

Schlechtweg, D., Eckmann, S., Santus, E., Schulte im Walde, S., & Hole, D. (2017). German in flux: Detecting metaphoric change via word entropy. In *Proceedings of the 21st Conference on Computational Natural Language Learning* (pp. 354–367). Vancouver, Canada.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*(2), 245–251.

Weeds, J., & Weir, D. (2003). A general framework for distributional similarity. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, Sapporo, Japan* (pp. 81–88).