An Inquiry into the Productivity of German Particle Verbs

Inna Stupak & Harald Baayen

Paper presented at the virtual Workshop on Morphology and Word Embeddings,
Tübingen/York, January 17, 2022

1 Introduction

This study addresses the productivity of German complex verbs. A series of studies have proposed criteria for establishing whether a derivational rule is productive, and what factors co-determine the productivity of a derivational rule (Schultink, 1961; Baayen, 2005; Plag, 1999; Dressler & Ladányi, 2000; Dressler, 2003; Fernández-Domínguez, 2009).

However, the statistical perspective of Baayen (2005) on derivational productivity does not address the question of what factors favor higher degrees of productivity. This brings us to the central problem addressed in this study: how can a word formation process, serving the function of providing names for new ideas and concepts, be productive, given that these concepts and ideas are themselves not predictable or compositional? If a word formation process gives rise to names that are unavoidably to a greater or lesser extent semantically idiosyncratic, how can it be productive at the same time?

In this study, we address the productivity and semantic transparency for German particle verbs (e.g., *backen* 'bake', *durchbacken*, 'bake through'; *Gift* 'poison', *vergiften* 'poison'; *kurz*, 'short', *abkürzen*, 'shorten'). German particle verbs have the property that the particle is separated from its base word when the verb occupies the initial position in the sentence, when the past participle inflects with *ge-*, and in infinitive clauses, in which *zu* appears in between particle and verb.

On the one hand, the meaning of some German particle verbs can be predicted given the particle and the verb, as in *retweeten* 'spread a tweet' (`https://www.owid.de/artikel/404004`), where *re-* means 'to do once more' and *tweeten* 'to tweet'. On the other hand, there are novel particle verbs whose meanings are not compositional. For example, the meaning of *downlocken* 'political decommissioning and suspension of economic and social activities (e.g. to contain an epidemic)' (`https://www.owid.de/docs/neo/listen/corona.jsp#`), cannot straightforwardly traced back to the meanings of *down* and *locken* 'to entice; to curl'.

A study of Mandarin adjective-noun compounds using distributional semantics (Shen & Baayen, 2020) reported that the category-conditioned degree of productivity varies with semantic transparency: the tighter the meaning relation between adjective and compound is, the greater the degree of productivity of an adjective will be. The present study follows up on the study of Shen & Baayen (see for related research, Bonami & Paperno, 2018) and applies their approach to German particle verbs, and follows this up with an inspection of the clustering of particle verbs in semantic space.

2 Productivity

Figure 1 plots, on a log-log scale, category-conditioned productivity against the number of types, for a total of 98 particles. Particles that have given rise to more types tend to be less productive. A Gaussian location-scale GAM (Wood, 2017) clarifies that the trend is nonlinear, and the variance decreases for increasing number of types. Thus, as observed by Shen & Baayen (2021) for Mandarin adjective-noun compounds, an increase in 'realized productivity' is detrimental for 'potential productivity'.

Figure 2 presents a regression model for 48 particles for which we could obtain sufficient numbers of word embeddings from fasttext (`https://dl.fbaipublicfiles.com/fasttext/vectors-crawl/cc.de.`

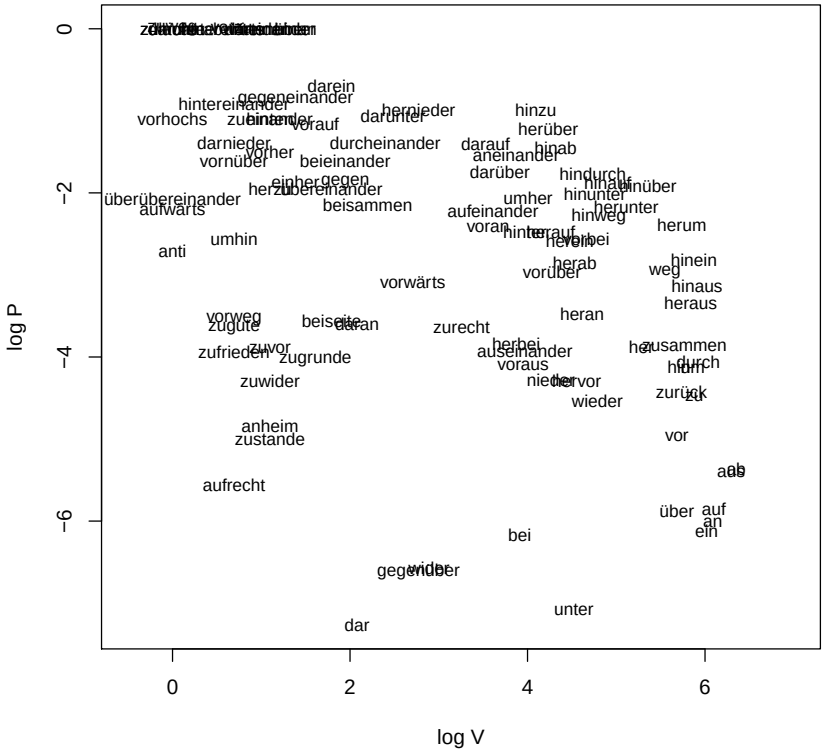

Figure 1: Category-conditioned productivity as a function of number of types, on a log-log scale: Productivity decreases as the number of types increases.

300.vec.gz) to calculate the mean, for each particle, of the correlations of the embedding of the particle and the embeddings of each of the particle verbs. This average correlation is a measure of semantic transparency: the more transparent a particle is, the stronger its mean correlation will be. A Gaussian location scale GAM indicated that log category-conditioned productivity (log P) increases with this average correlation, whereas the variance in log P decreases with increasing correlation. The color coding in the plot indicates the number of semantic functions (such as locative or temporal) listed for the particle in red German dictionaries, which ranged from 1 (black), two (red) to more than two (blue). Thus, particles that realize more different semantic functions tend to be less productive. Thus, both a measure grounded in distributional semantics, and a classical lexicographic measure, support the hypothesis that 'potential' productivity should decrease with decreasing semantic transparency.

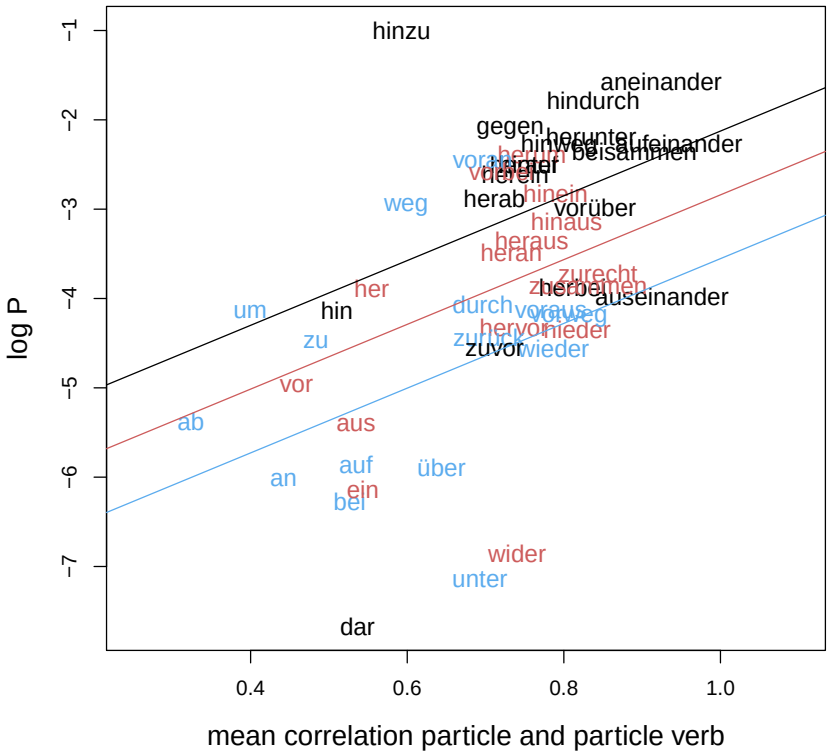

Figure 2: Category-conditioned productivity as a function of the mean correlation of the semantic vector of the particle and the semantic vectors of the particle verb. Color coding: black lines and words describe words that have one semantic function (e.g., locative, temporal), red lines and words concern words with two semantic functions. Blue lines and words pertain to words with more than two semantic functions.

3 Exploring the semantics of particle verbs with t-SNE.

T-SNE is a dimension reduction technique (van der Maaten & Hinton, 2008) that is optimized for preserving in a two-dimensional plane the distances between points in a high-dimensional space. If there are clusters in a high-dimensional space, t-SNE should be able to find these with high probability. Figure 3 shows that only very few clusters are detected for 22 particles with at least 10 formations in our dataset.

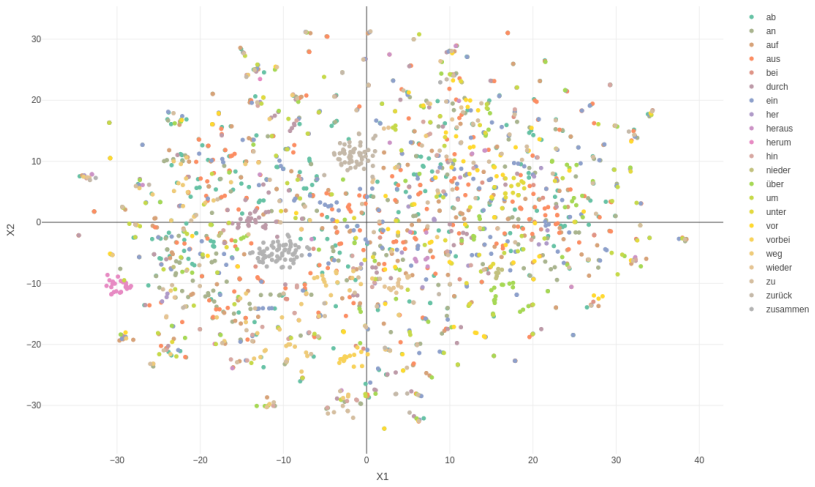

Figure 3: Locations of words with 22 particles in t-SNE space. Only a few clusters are visible: *zurück* (light brown, upper center), *zusammen* (grey, to the lower left of the origin), *herum* (pink, far left), and *durch* (dark pink, slightly above zusammen.

Figure 4: Locations of the shift vectors for words with 22 particles, in t-SNE space. The clusters at the outer periphery bring together formations with different particles but the same verb stem.

The only clusters that emerge are for *zurück* (light brown, upper center), *zusammen* (grey, to the lower left of the origin), *herum* (pink, far left), and perhaps *durch* (dark pink, slightly above zusammen).

When we consider the shift vectors, we find some clustering, but this clustering is by verb stem, rather than by particle, as shown in Figure 4. This suggests to us that for these verbs, the semantic contribution of the different particles are apparently very similar.

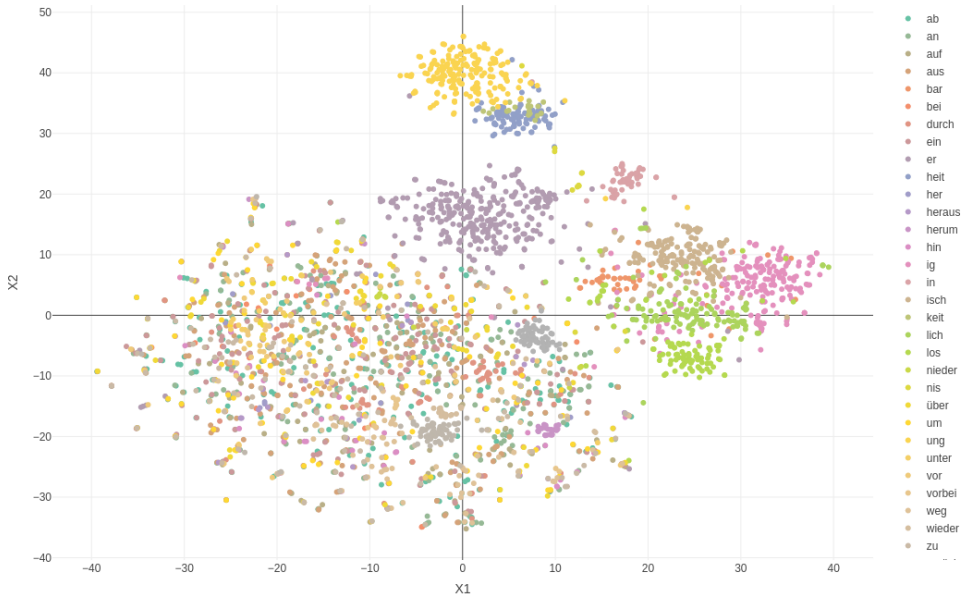

Figure 5: Locations of the shift vectors for particle verbs and derivational suffixes in t-SNE space. The particle verbs are found in the lower left, adjectives in the center right, and nouns in the upper center. Words cluster by their derivational exponents, but hardly ever by their particles.

To better understand the patterning of the particle verbs, we ran a t-SNE clustering on the particle verbs and 11 German derivational suffixes (*bar, er, heit, ig, in, isch, keit, lich, los, nis*, and *ung*). Figure 5 clarifies that words with the derivational suffixes cluster by suffix with remarkably little overlap. But for the set of suffixes, there is no systematic relation between number of types and P, as illustrated in Figure 6 ($r = -0.27, p = 0.43$). In summary, we observe the following:

1. The particles are characterized by a negative correlation between P and V, but they do not cluster in t-SNE maps.

2. Derivational suffixes do not show a negative correlation, but they cluster beautifully in t-SNE maps.

4 Discussion

The dissociation of clustering and V-P correlations (and probably, no correlation of P and transparency, but we need to test this) may have several causes. One possibility is that we are observing a difference in headedness, but some particles can change word category. Another possibility is that we have a confound between prefixation and suffixation. A third factor to take into account is that the particles can have many senses (e.g., 18 senses for *ab*, and 40 for *an*), whereas the suffixes tend to be semantically much more constrained. For instance, *auffallen* (to be striking, salient) and *ausfallen* (to fail) have very different meanings, which are different from the meanings of *fallen* (to fall), *auf* (on, up), and *aus* (from). *Absolutheit* (absoluteness) and *Alleinheit* (loneliness), by contrast, are straightforward nominalizations of the adjectives 'absolute' and 'alone'. Furthermore, it is worth noting that for particle verbs, the word embeddings are

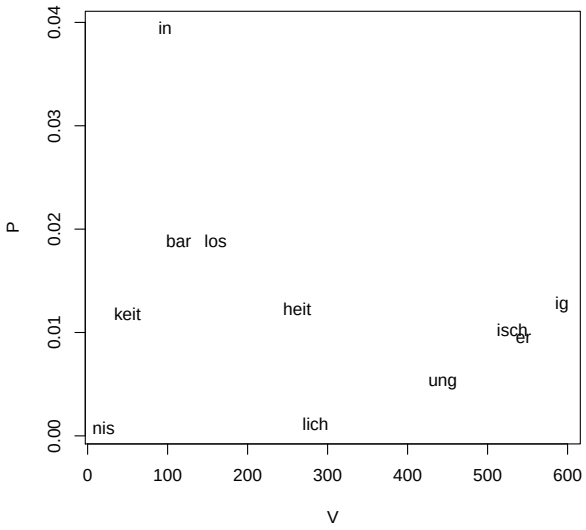

Figure 6: V and P are uncorrelated for the derivational suffixes.

likely too imprecise due to the particle not being concatenated with the verb stem in many occurrences. Fourth, many particles are used independently, and frequently, as prepositions, so the word embeddings for the particles may be confounded with prepositional semantics rather than the often more aspect-like or metaphorical semantics of the particles.

However, if the present results are catching a glimmer of the truth, then it is conceivable that particle verbs in German are more similar to compounding in Mandarin than to suffixation in German. It is worth noting that both constituents in Mandarin compounds and German particle verbs have their own constituent families, which in the case of particles include (due to the imperfections of our embeddings) prepositional families (see Baayen et al., 2011, for a discussion of prepositional entropy). Suffixes, on the other hand, are combinatorially much more restricted. We think that it is precisely the fact that suffixes do not combine to the right in the way that left constituents of compounds, and the particles of German particle verbs, do, which makes it possible for suffixes to have the highly distinct semantic effects that are visible in the t-SNE maps.

References

Baayen, R. H. (2005). Corpus linguistics in morphology: morphological productivity. In Lüdeling, A., Kytö, M., and McEnery, T., editors, *Handbook of Corpus Linguistics (Handbücher zur Sprach- und Kommunikationswissenschaft)*, page in press. De Gruyter.

Baayen, R. H., Milin, P., Filipović Durdević, D., Hendrix, P., and Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, 118:438–482.

Bonami, O. and Paperno, D. (2018). Inflection vs. derivation in a distributional vector space. *Lingue e Linguaggio*, 17(2):173–195.

Domínguez, J. F. (2009). *Productivity in English Word-formation: An Approach to N+ N Compounding*, volume 341. Peter Lang.

Dressler, W. U. (2003). Degrees of grammatical productivity in inflectional morphology. *Italian Journal of Linguistics*, 15:31–62.

Dressler, W. U. and Ladányi, M. (2000). Productivity in word formation (wf): a morphological approach. *Acta Linguistica Hungarica*, 47(1):103–145.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Plag, I. (1999). *Morphological productivity: structural constraints in English*. Mouton de Gruyter, Berlin.

Schultink, H. (1961). Produktiviteit als morfologisch fenomeen. *Forum der Letteren*, 2:110–125.

Shen, T. and Baayen, R. H. (2021). Adjective–noun compounds in mandarin: a study on productivity. *Corpus Linguistics and Linguistic Theory*.

Wood, S. N. (2017). *Generalized Additive Models*. Chapman & Hall/CRC, New York.