Assessing the diachronic change of a word-formation pattern: Frequency, productivity, and interaction patterns

Stefan Hartmann (University of Hamburg, Germany)

Nominalization with the suffix *-ung* (e.g. *Landung* 'landing', *Bildung* 'education') is certainly among the most well-studied word-formation patterns in German, both synchronically and diachronically (see Hartmann 2016 for a recent overview). Most importantly from a historical-linguistic perspective, Demske's (2000) corpus-based study of *ung*-nominalization in the Early New High German period (1350–1650) has shown that the pattern, while becoming more frequent, suffers a considerable decrease in morphological productivity. In addition, she argues that the word-formation pattern becomes more "nominal" over time: More and more *ung*-nominals denote concrete objects (e.g. *Heizung* 'heating device') or even persons (*Bedienung* 'waiter/waitress). While much of Demske's (2000) study remains qualitative, Hartmann (2016) has conducted a systematic quantitative corpus study using larger corpora and extending the scope of investigation from the Early New High German period to the beginning of the New High German period (from 1650 onwards). Using the Mainz Early New High German Corpus (Kopf 2016) and the GerManC corpus (Durrell et al. 2007), it could be shown that *ung*-nominals, over time, are increasingly attracted to prototypically "nominal" constructions (e.g. determiner constructions and plural constructions), while their frequency in constructions that evoke a verb-like construal drops significantly.

One drawback of Hartmann's (2016) study, however, is that the corpora are fairly small and only comparable to a limited extent. In addition, the study suggests that the relevant changes are still in progress at the end of the time period covered by the GerManC corpus, which is why it seems promising to investigate the pattern in a corpus that covers the 19th century as well. Therefore, the present study replicates the results obtained by Hartmann (2016) using the German Text Archive (*Deutsches Textarchiv*, DTA). The DTA is a 100-million-word corpus covering the time span from 1600 to 1900. As the DTA is very unbalanced both for time periods and for text types, a smaller subcorpus (DTA*baby*) has been compiled comprising 270 texts of normalized length, balanced for fifty-year periods and three text types, and comprising about one million tokens. Fig. 1 shows an overview of the composition of both corpora.

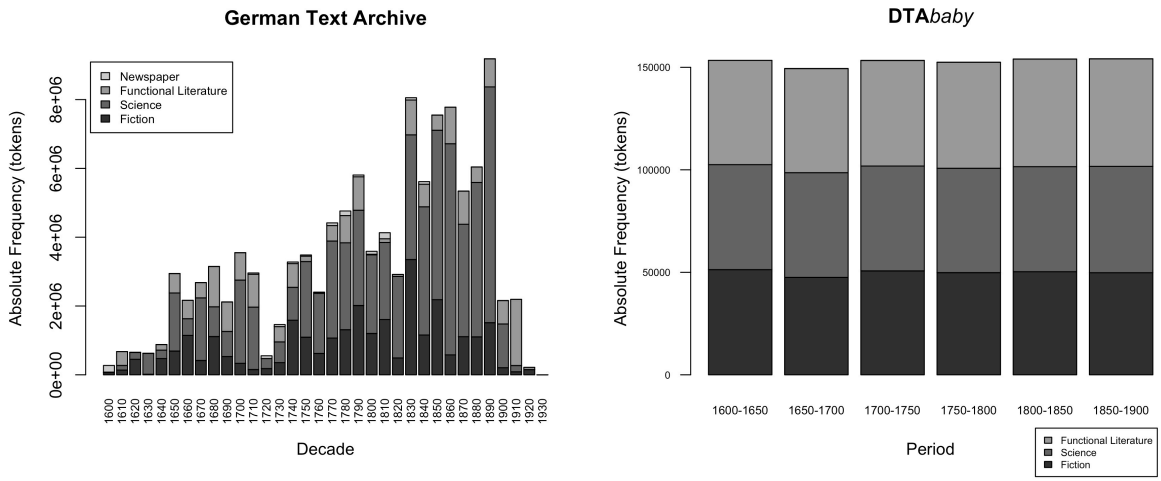

Fig. 1: Overview of the absolute token frequencies of DTA (left panel) and DTA*baby* (right panel). Note that newspaper texts have been omitted from DTA*baby* as they are heavily underrepresented in the DTA data.

	Tokens	Types	Hapax Legomena
Period 1 (1600-1649)	922	335	84
Period 2 (1650-1699)	924	384	93
Period 3 (1700-1749)	1273	395	69
Period 4 (1750-1799)	2106	501	64
Period 5 (1800-1849)	2720	614	102
Period 6 (1850-1899)	3001	663	159
Sum	10946	2892	571

Tab. 1: ung-nominalization in the DTA*baby* corpus: Overview of the corpus data.

Using the DTA*baby* data, it can be shown that *ung*-nominalization experiences a steep increase in token frequency throughout the entire period covered by the corpus, as Tab. 1 shows. Using Baayen's (e.g. 2009) measure of potential productivity, i.e. the ratio of hapax legomena to the total number of items belonging to the construction in question, suggests that the productivity of the pattern decreases until the end of the 18th century but then sees a slight increase. However, given the significant differences in token frequency, comparing the potential productivity values of the individual corpus periods is highly problematic (see e.g. Gaeta & Ricca 2006; Säily 2011, among many others). For this reason, a finite Zipf-Mandelbrot model (see e.g. Baayen 2001; Evert 2004) was used for extrapolating the number of hapaxes that can be expected for an arbitrarily large total number of tokens. As the Zipf-Mandelbrot model, unlike Baayen's productivity measures, does not require equal sample sizes, all 1,713,147 *ung*-nouns attested in the complete DTA corpus were used for obtaining the extrapolations. As the goodness-of-fit of the resulting model proved suboptimal, a bootstrapping approach was used in addition to the simple model. For each of the three centuries covered by the DTA, 100,000 attestations were randomly sampled, and a Zipf-Mandelbroot model was fit to the data using *zipfR* (Evert & Baroni 2007). This procedure was repeated 100 times. The left panel of Fig. 2 shows the results. What appears, in the plot, as a thick dark-grey line consists of 100 individual lines that represent the fZM models fit to the 17th century data. The same goes for the (partly overlapping) areas that appear in somewhat lighter shades of grey which represent the 18th and 19th century data, respectively. The black lines represent the average growth curve, obtained by calculating the mean *V1* (= number of hapaxes) for each *N* (= token frequency). The right panel shows the extrapolated potential productivity for an arbitrary value of *N*=500,000.

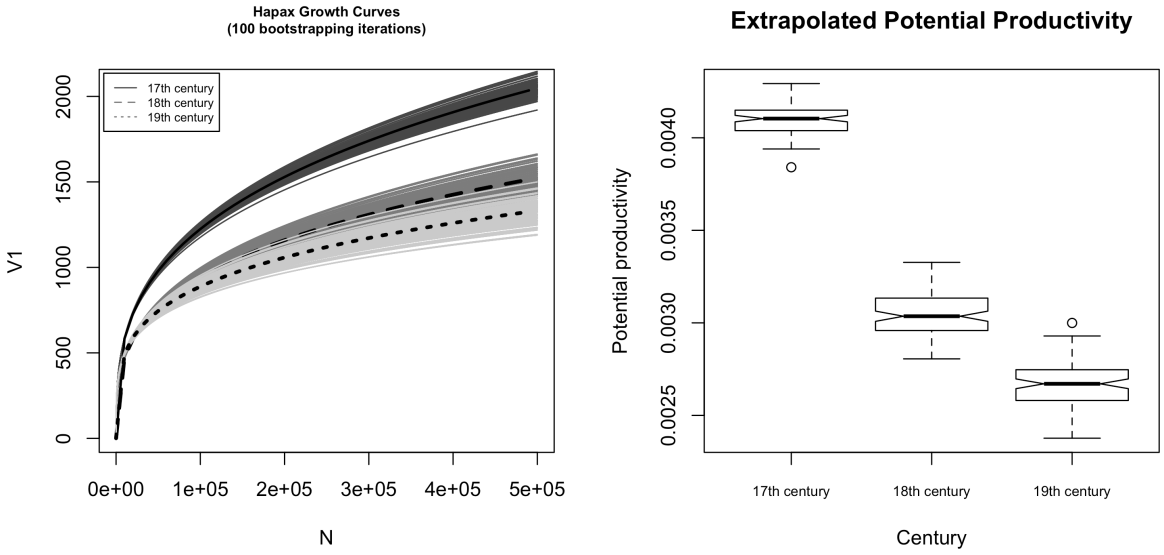

Fig. 2: Left panel: Hapax growth curves for random samples of 100,000 tokens per century. Right panel: Extrapolated productivity for N=500,000.

In sum, the extrapolated values suggest that the morphological productivity of the pattern continues to decrease, as observed by Demske (2000) and Hartmann (2016) for other corpora. While the decrease is clearly observable from the 17th to the 18th century, the picture for the 18th/19th century is not entirely clear – while some models suggest a slight increase in potential productivity, others suggest that it keeps decreasing. Nevertheless, the overall picture is very clear and confirms the findings from previous literature that all in all, the morphological productivity of *ung*-nominalization wanes throughout the New High German period.

However, equally important for understanding the diachronic developments of a word-formation pattern are what can be called "interaction patterns" from a Construction Grammar perspective, according to which more abstract morphological or syntactic patterns are considered form-meaning pairs (constructions) in their own right (cf. e.g. Goldberg 2006). Demske (2000) and Hartmann (2016) have already shown that the pattern's constructional preferences change significantly over time, reflecting its drift towards a higher degree of "nouniness". Three constructions are particularly interesting in this regard: The determiner construction, the plural construction, and what can be called the [P NOM] construction, in which a nominalization is used as the complement of a preposition (without a determiner). The latter is particularly interesting in that it tends to evoke a highly processual construal of the nominal in question. Consider, for instance, *grabung* 'digging' in (1) (from Demske 2000: 380).

(1) *Diese wochen hat man alhie **in grabung** deß Grunds zu S. Petro ein Kreutzlein oder heyligthumb* [...] *gefunden.* 'This week, **in digging** the ground of St Peter's [cathedral], a cross or sanctuary has been found.' (Relation des Jahres 1609)

The proportion of *ung*-nominals in the [P NOM] construction decreases constantly throughout the period covered by the DTA*baby* corpus (Kendall's τ=-1, T=0, p<0.01; see Fig. 3).

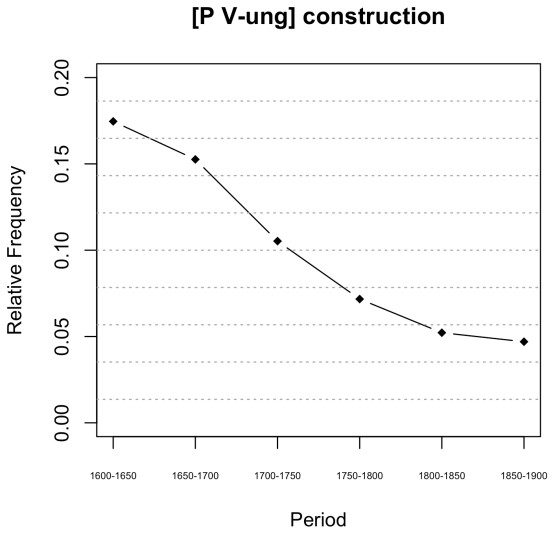

Fig. 3: Relative frequency of ung-nominals in [PREP NOM] constructions in relation to the total number of ung-nominals in the respective corpus period as attested in the DTA*baby* corpus.

Conversely, the proportion of *ung*-nominals with a determiner and the proportion of pluralized *ung*-nominals continue their increase that started in the Early New High German period (see e.g. Demske 2000). Both the use of determiners and pluralization can be seen as diagnostics of increasing "nouniness" (see e.g. Vogel 1996; Fonteyn 2016; Bekaert & Enghels forthc.). The proportion of *ung*-nominals with a determiner increases slightly, but significantly (Kendall's τ=1, T=15, p<0.01). In the case of pluralization, the initial increase is followed by a slight decrease, but altogether, it is fairly clear that we find more pluralized instances in later than in earlier corpus periods.

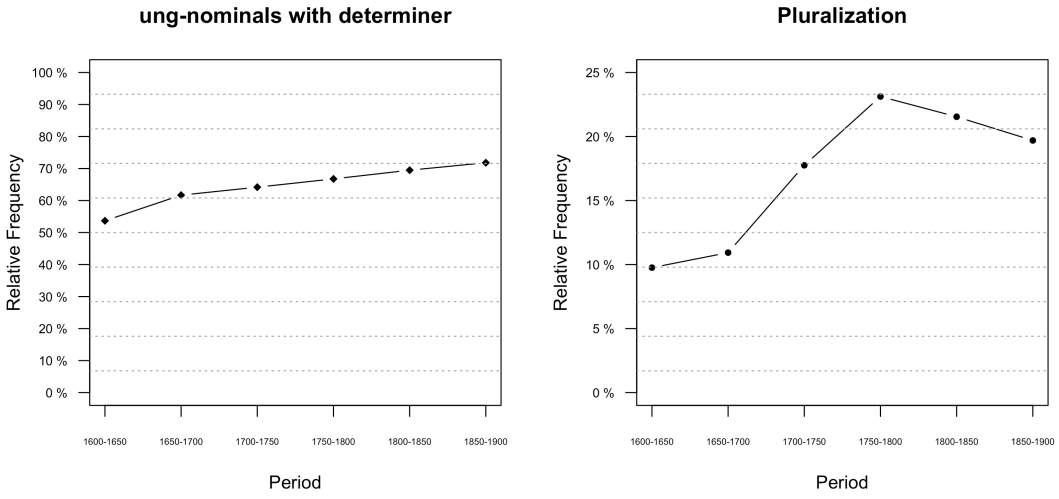

Fig. 4: Relative frequency of ung-nominals with a determiner (left panel) and of pluralized ung-nominals (right panel) in relation to the total number of ung-nominals in the respective corpus period.

All in all, then, the data from the DTA corpus and from the newly-compiled sample corpus DTA*baby* lend further support to the hypothesis that the diachrony of *ung*-nominalization can be described as a "nominalization process with 'nominalization' taken literally" (Demske 2002: 69). Such processes of "nominalization" can be observed in other cases, and in other languages, as well. For instance, Fonteyn & Hartmann (2016) have shown that English *ing*-nominals undergo a surprisingly similar development. But while the tendency of abstract nouns to develop more concrete meanings – posited by e.g. Panagl (1987: 146) as a widespread cross-linguistic tendency – has so far been an observation largely based on qualitative analysis of the available data, corpus-linguistic studies like the one presented here can help empirically substantiate such hypotheses drawing on in-depth quantitative analyses of individual word-formation patterns.

References

Baayen, R. H. (2001). *Word Frequency Distributions*. Dordrecht: Kluwer.

Baayen, R. H. (2009). Corpus Linguistics in Morphology: Morphological Productivity. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics* (pp. 899–919). Berlin, New York: De Gruyter.

Bekaert, Elisa & Renata Enghels. (forthc.) On the edge between nouns and verbs. The heterogeneous behavior of Spanish deverbal nominalizations empirically verified. To appear in *Language Sciences*.

Demske, U. (2000). Zur Geschichte der ung-Nominalisierung im Deutschen: Ein Wandel morphologischer Produktivität. *Beiträge Zur Geschichte Der Deutschen Sprache Und Literatur*, *122*, 365–411.

Demske, U. (2002). Nominalization and Argument Structure in Early New High German. In E. Lang & I. Zimmermann (Eds.), *Nominalisations* (pp. 67–90). Berlin: ZAS.

Deutsches Textarchiv (DTA). http://deutschestextarchiv.de/

Durrell, M., Ensslin, A., & Bennett, P. (2007). The GerManC Project. *Sprache Und Datenverarbeitung*, *31*, 71–80.

Evert, S. (2004). A simple LNRE model for random character sequences. In *Proceedings of JADT 2004* (pp. 411–422).

Evert, S., & Baroni, M. (2007). zipfR: Word frequency distributions in R. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Posters and Demonstrations Sessions, Prague* (pp. 29–32).

Fonteyn, L. (2016). *Categoriality in language change. The case of the English gerund*. PhD Thesis, KU Leuven.

Fonteyn, L., & Hartmann, S. (2016). Usage-based perspectives on diachronic morphology: A mixed-methods approach towards English ing-nominals. *Linguistics Vanguard*, *2*(1). http://doi.org/10.1515/lingvan-2016-0057

Gaeta, L., & Ricca, D. (2006). Productivity in Italian word-formation. *Linguistics*, *44*(1), 57–89.

Goldberg, A. E. (2006). *Constructions at Work: The Nature of Generalization in Language*. Oxford: Oxford University Press.

Hartmann, S. (2016). *Wortbildungswandel. Eine diachrone Studie zu deutschen Nominalisierungsmustern*. Berlin, Boston: De Gruyter.

Kopf, Kristin. (2016.) Mainzer (Früh-)Neuhochdeutschkorpus, 1500–1720. Mainz.

Panagl, O. (1987). Productivity and Diachronic Change in Morphology. In W. U. Dressler (Ed.), *Leitmotifs in Natural Morphology* (pp. 127–151). Amsterdam and Philadelphia: John Benjamins.

Säily, T. (2011). Variation in morphological productivity in the BNC: Sociolinguistic and methodological considerations. *Corpus Linguistics and Linguistic Theory*, *7*(1), 119–141. http://doi.org/10.1515/cllt.2011.006

Vogel, P. M. (1996). *Wortarten und Wortartenwechsel: Zur Konversion und verwandten Erscheinungen im Deutschen und in anderen Sprachen*. Berlin, New York: De Gruyter.