Quantifying regularity in morphological processes: An ongoing study on nominalization in German

Rossella Varvara\textsuperscript{1}, Gabriella Lapesa\textsuperscript{2}, Sebastian Padó\textsuperscript{2}

University of Trento, Center for Mind and Brain Sciences\textsuperscript{1}
University of Stuttgart, Institute for Natural Language Processing\textsuperscript{2}

Nominalization is a highly productive process of derivation in many languages, and often there are multiple nominalization patterns for a verbal base that are associated with (more or less) subtly different semantics. In German, two nominalization patterns are particularly prominent in that they are applicable to a broad range of verbal bases: nominal infinitives (NI, *evaluieren*, “to evaluate” → *das Evaluieren*, “the evaluating”) and -ung deverbal nouns (UNG, *evaluieren*, “to evaluate” → *die Evaluierung*, “the evaluation”). At a general level, the semantics of UNGs and NIs can be considered quite similar. This is not only because they are transpositional, i.e., they change the part-of-speech of the base term, but also because of the tight relation between the action/event denoted by the base verb and the corresponding nominalizations.

The study presented in this abstract proposes to investigate the distinction between UNGs and NIs with methods from distributional semantics that capture the meaning shifts associated with the two nominalization patterns. We hypothesize that despite the shared syntactic features, the meaning shifts of UNGs and NIs are markedly different on the distributional level.

Furthermore, we place the comparison between the two nominalization processes within the broader theoretical perspective of a graded distinction between inflection and derivation. In particular, we hypothesize that while the distributional behavior of UNGs places them close to the derivation extreme of the derivation/inflection continuum, NIs pattern with the cases of inflection. To support this claim, we compare the nominalizations to two landmarks: a clear case of inflection, namely present participles (*evaluierend*, “evaluating”) and an unequivocal case of derivation, namely agent nouns (*der Evaluierer*, “the evaluator”).

**Theoretical background.** It has been frequently argued that the main difference between inflectional and derivational morphology is one of regularity (Haspelmath 1996, Laca 2001, among others): inflection is more regular than derivation with respect to both productivity and semantic shifts. In this work, we focus on the latter aspect and attempt to operationalize the often rather vague concepts from the literature using the concepts of transparency and specificity.

**Transparency** is quantified in terms of semantic similarity between the input (base) and the output (derived or inflected word) of a morphological process, and it can also be interpreted in terms of compositionality. It is commonly assumed that derivatives, differently from inflected forms, tend to be only partly compositional and transparent with respect to their base because of the stronger semantic shift they produce (Booij 2000:364, Laca 2001:1217). Within the nominalization domain, in which verbs (denoting events/states) are turned into nouns (which may or not denote an event/state), the base/derived similarity can be effected by type shifts (when the derivative denotes an entity) and by type ambiguity (when the derivative denotes an event or entity).

**Specificity** quantifies the information content that is added in a nominalization process when the new word is created; according to the Monotonicity Hypothesis (Koontz-Garboden, 2007), word formation operations may or may not add information to the semantic representation of the base, but never remove information. Adopting this stance, we operationalize the notion of specificity as a binary feature: unmarked, when the derivative keeps the event reading of the base; marked, when additional meaning is added (e.g., to create an agent noun).

We predict that the two nominalization processes introduced above differ in terms of transparency/similarity and specificity/markedness. The left panel of Figure 1 displays our predictions concerning the location of UNGs and NIs relative to each other and to the landmarks PPs and ERs as a function of transparency and specificity:

- NIs are unmarked and undergo type shift only in idiomatic uses (see discussion below); they pattern with PPs, our inflection landmark;
- UNGs are marked and less similar than NIs to their bases, due to semantic type ambiguity (*die Evaluierung* can refer both to the action of evaluating, and to its result).

As shown in the prediction diagram, we do expect a certain degree of category overlap. We foresee that at least three cases will make our picture less tidy: (1), lexicalized UNGs (e.g., *Vorlesung*, “the lecture”, derived from *vorslesen*, “to read aloud”), which show a major change in meaning; (2), idiomatic NIs (e.g., *das Essen*, “the
food” instead of “the eating”), which are more specific and less transparent with respect to their bases due to a semantic type shift (das Essen refers to an object and not to an action, like other NIs); and (3), semelfactive ERs (e.g., Hüpfer, “a jump”, derived from Hüpfen, “to jump”), which, differently from other ERs which refer to individuals (agents or instruments), refer to events and are therefore unmarked and more similar to their bases.

Experimental Setup. The experimental items for our dataset are sampled from the lemmatized SdeWaC corpus (Faaß and Eckart, 2013). Quadruples of NIs, UNGs, ERs, and PPs for the same base verb are extracted from the DErivBase (Zeller et al. 2013), a large-coverage database of German derivational morphology. For our experiment, we keep only quadruples for which all items have a frequency value in the middle range (two standard deviations) of the frequency distribution of the respective class (nouns, for NIs, -ungDNs, -erDNs; adjectives, for PPs). The sampling procedure results in a dataset containing 115 quadruples.

We employ Distributional Semantics Models (Miller and Charles, 1991) to ground the theoretical predictions outlined above. For each word in the dataset (derived and base terms), distributional count vectors have been extracted from the SdeWaC corpus (Faaß and Eckart, 2013), adopting a symmetric 5-words context window (10k most frequent content words as context dimensions) with PPMI as a scoring function.

We quantify the transparency of the derivative with respect to the base in terms of feature inclusion (InvCL, Lenci and Benotto 2012). Inclusion measures can be considered a more fine-grained, asymmetric counterpart of standard distance metrics (such as cosine or Euclidean distance) because they quantify the extent to which the features of the derived word form a subset of those of the base word.

Specificity finds its distributional correlate in the nearest-neighbor density measure, computed as the average distance between the top nearest neighbors of the target word and the word itself: the denser the neighborhood, the more specific the meaning of the target word (Marelli & Baroni 2015; Sagi et al. 2009). More specifically, we consider the difference between the density of the base and of the derivative.

Results. The right-hand panel of Figure 1 shows a scatterplot of derived words, using transparency (InvCL) and specificity (Δ density) as coordinates. Despite the obviously large variance inherent in the empirical data, the corpus-driven model bears out the main theoretical hypotheses: (a), NIs and PPs group together in the top left corner, with high transparency and low specificity; (b), UNG lives in the top right corner, with still high transparency but higher specificity; (c) ERs exhibit a substantially lower transparency than all other categories. However, we also see some unexpected outcomes, notably (d) the relatively low specificity of many ER derivatives, and (e) the difficulty of distinguishing PPs and NIs at all. We are in the process of further investigating (d) and (e) and are carrying out a more rigorous statistical analysis of the predictions as well as a qualitative analysis of the many outliers we see for all classes to obtain a more comprehensive understanding of the patterns.

1We also conducted experiments based on vector entropy as an alternative to density (Santus et al. (2014); Padó et al. (2015)); the results showed the same trends of the neighbor density ones, and are not reported here for reasons of space.
References


