Applying Morphological Productivity Measures to Syntactic Constructions: German Comparatives and the je... destino Constructions

Amir Zeldes (amir.zeldes@rz.hu-berlin.de) – Korpuslinguistik / SFB 632 Teilprojekt D1
Institut für deutsche Sprache und Linguistik, Humboldt-Universität zu Berlin

Morphological Productivity

- **Productivity** has been addressed mainly in morphology as the binary ability of a word formation process to produce *new words* or else the scalar degree of how easily new items arise in that formation (Bauer 2001).
- Some definitions concentrate on spontaneous generation of items not encountered before by the speaker through regular combination of a compatible base with a word formation process to produce a transparent item whose meaning can be inferred from the base and the formation, e.g. a stem and a suffix: 
  
  miniaturisier*, trans-stem + bar*adj-bar* → miniaturisierbar *capable of being miniaturized*

- The degree of productivity is often associated with *type frequency* of the formation (i.e. how many adjectives with -bar are there?), which can be measured in a corpus, and the proportion of productive cases therein.
- It is difficult to determine for all items whether the speaker was familiar with them and whether they are transparent.
- Baayen (2001, 2009) uses *hapax legomena*, words appearing only once in a corpus, to estimate productivity. The reasoning is that neologisms form a subset of these, though words appearing two or three times may also be relevant through repetition of a neologism: neologisms ≤ hapax legomena (U d is r's legomena).

Baayen’s Productivity Measures

Baayen defines for a corpus of N words and word formation process C:
- \( V(C, N) \) is the total count of tokens from C in N
- \( V(N, C) \) is the count of type distinct items from C in N
- \( V(1, N) \) is the total amount of hapax legomena in N
- \( V(1, C, N) \) is the type count of distinct items from C appearing once in N

From these data he derives three productivity measures for C in N:

1. **Extent of Use** = \( \frac{V(C, N)}{N} \)
   - corresponds to productivity of C in the language up till now – how many types has it created?
2. **Hapax-conditioned Degree of Productivity** = \( \frac{V(C, N)}{V(N, C)} \)
   - corresponds to expanding productivity – what portion of the hapax does C contribute?
3. **Category-conditioned Degree of Productivity** = \( \frac{V(C, N)}{V(1, C, N)} \)
   - corresponds to saturation of C or how likely it is to produce more words in the future – what proportion of tokens in C are hapax legomena?

Measuring the Productivity of German Comparatives

Using Baayen’s measures and the frequencies of all comparatives in a corpus we can compute productivity, for example for German comparatives derived from adjective bases with the suffix -er.

\( N = c’t-Magazin + Parlamentsreden + EuroParl: 14 + 37 + 27 = 78 M \) Token

- **Type Frequency** of je... destino:
  - je COMP shows more variety than destino COMP, though unlike destino, it exhibits a smaller spectrum than statistically predictable limited semantics?
  - The verbless CC is especially limited: je COMP (…) destino COMP is mostly restricted by three types of, which becomes ~ 73%, though it never follows je. Only two hapax outside core vocabulary: ergonomischer* more ergonomic* and hilfloser* more helpless*. The complement CC has very different types e.g. besser is attested after je. Verbless CC’s have different usage (more lexicalized?)

- **Extent of Use** shows unsurprisingly that je... destino COMP are very rare uses of the comparative, and je... destino COMP very rarely manifests itself

- **Hapax productivity** shows je and destino are responsible for little productivity in comparatives, but their category productivity shows they have the potential for many novel constructions.