Applying Morphological Productivity Measures to Syntactic Constructions: German Comparatives and the *je … desto* 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$_{\text{v.trans.-stem}}$ + *bar*$_{\text{adj-suf}}$ → *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 dis/tris legomena)

Baayen's Productivity Measures

Baayen defines for a corpus of N words and word formation process C:

- N(C) is the token count from C in N
- V(C,N) is the type count of 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 = $V(C,N)$

   corresponds to productivity of C in the language up till now – how many types has it created?
2. Hapax-conditioned Degree of Productivity = $\frac{V(1,C,N)}{V(1,N)}$

   corresponds to expanding productivity – what portion of the hapax in the corpus does C contribute?
3. Category-conditioned Degree of Productivity = $\frac{V(1,C,N)}{N(C)}$

   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)

•Extent of Use = V(comp, 78,637,399) = 1969

•Hapax-conditioned = 780/565020 =.00138

•Category-conditioned = 780/113196 = .00689

type	freq
besser	18270
später	7983
stärker	6844
…	
ökoverträglicher	1
objektorientierter	1
niedlicher	1
nobler	1
notebook-freundlicher	1

Comparing this to productivity measures of other processes gives an intuitive idea of the meaning of these results:

	-ung nouns	comparative	superlative
Extent	43433	1969	1494
Hapax	0.043639	0.00138	0.001215
Category	0.015611	0.00689	0.011497

Literature:
Baayen, R. H. 2001. *Word Frequency Distributions*. Dordrecht / Boston / London: Kluwer Academic Publishers.
Baayen, R. H. 2009. Corpus Linguistics in Morphology: Morphological Productivity. In: Lüdeling, A. & Kytö, M. (eds.), *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter, 899-919.
Bauer, L. 2001. *Morphological Productivity*. (Cambridge Studies in Linguistics 95.) Cambridge: CUP.
Beck, S. 1997. On the Semantics of Comparative Conditionals. *Linguistics and Philosophy* 20, 229-271.
Culicover, P.W. & Jackendoff, R. 1999. The View from the Periphery: The English Comparative Correlative. *Linguistic Inquiry* 30(4), 543-571.
den Dikken, M. 2005. Comparative Correlatives Comparatively. *Linguistic Inquiry* 36(4), 497-532.
Kiss, T. .2007. Produktivität und Idiomatizität von Präposition-Substantiv-Sequenzen. *ZS* 26(2), 317-345.
Goldberg, A. E. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: UCP.
Goldberg, A. E. 2006. *Constructions at Work: The Nature of Generalization in Language*. Oxford: OUP.
Zifonun, G., Hoffmann, L. & Strecker, B. 1997. *Grammatik der deutschen Sprache, 3*. Berlin etc.: de Gruyter.

Applying the Measures to *je…desto* Comparative Correlatives

- Syntactic constructions can be seen as similar to morphological formations:

  Regular formation & transparent meaning from constituents + construction (cf. Goldberg 1995, 2006)
- Comparative correlatives' constructional compositionality particularly called into question (Culicover & Jackendoff 1999, Beck 1997, den Dikken 2005)

Problems and Questions:

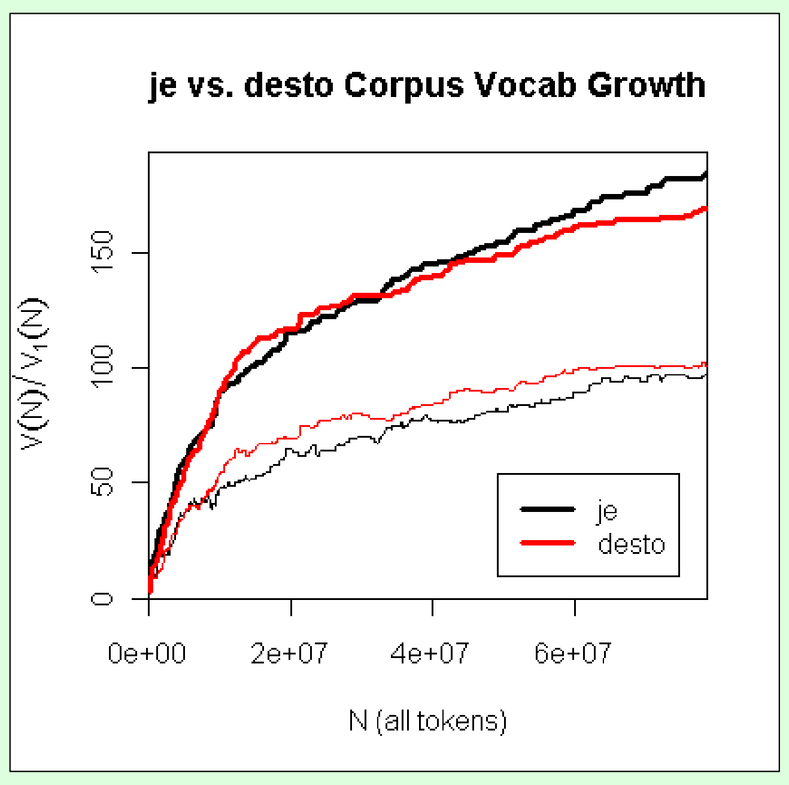
- What is **productivity for syntactic constructions**? (cf. Kiss 2007)
  → Attempt to use same criteria: unencountered, regular, transparent
- Can the productivity of constructions be **quantified** in a similar way?
  → **Empty positions** in construction determine type
  → Use '***hapax syntassomena***' to calculate measures
  → use n^th^ **root** of hapax count for multiple slots, or **average** of slot scores?
- **How can N be defined** for syntactic productivity?
  → Number of tokens as estimate for corpus size? Number of constructions?
  → How do we count how many constructions appear in a corpus in total?
  → For fixed length constructions: number of times it fits in the corpus?
  → Can we ignore N for same corpus comparisons? (measures not 0-1)

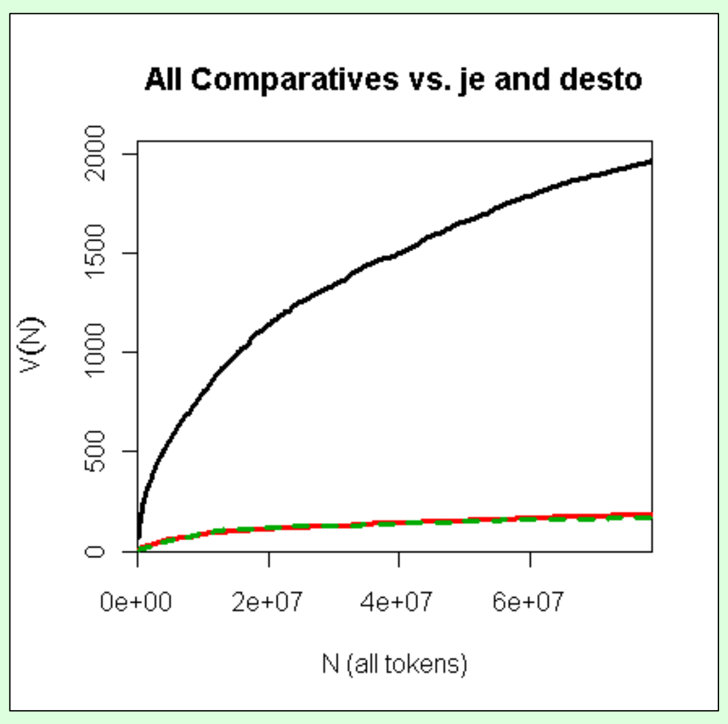
Constructional Predictions for Comparative Correlatives:

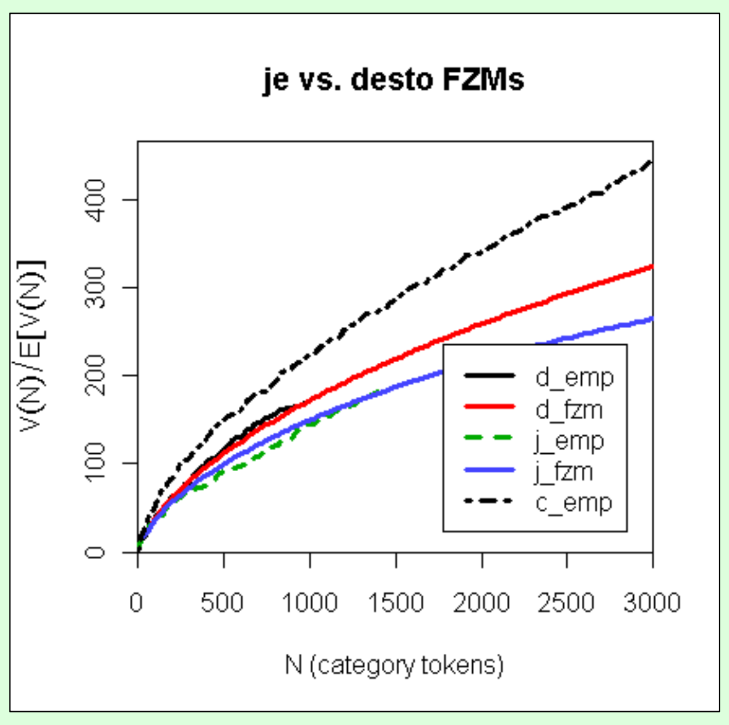
- As subsets of comparatives, CCs will trivially be more restricted
- However, they must be compatible with CC semantics, thus we expect even less type variability than statistically predicted by frequency alone
- Since *desto* is usually used to present the benefit correlated with some property, we expect a set of value-judging adjectives with little productivity
- Since *je* expresses the properties leading to these benefits, which can be more diverse, we expect more productivity, but still much less than expected from the pure productivity of comparatives
- Since constructions are form-meaning pairs, variants will show different lexical/productive behavior. Hence e.g. claims that verbless CC's are cases of copula ellipsis (Zifonun et al. 1997: 2338) should be falsifiable

Results

type	je	dest.	j X d	j d X	j s	d s	freq
besser	70	212	0	37	4	3	18270
später	22	5	0	0	0	0	7983
stärker	65	56	0	0	3	3	6844
ferner	0	0	0	0	0	0	5975
länger	179	23	2	0	2	0	4659
schneller	88	40	8	0	0	0	4423
lieber	0	1	0	1	0	0	4281
höher	179	76	0	1	9	7	3330
größer	195	120	4	2	5	11	3126
…							







- *je COMP* shows more variety than *desto COMP*, though unlike *desto*, it exhibits a smaller spectrum than statistically predictable→limited semantics?
- The verbless CC is especially limited: *je COMP (,) desto COMP*. *Desto* is followed by only 13 types, of which *besser* = 73%, though it never follows *je*. Only two hapaxes outside core vocabulary: *ergonomischer* 'more ergonomic' and *hilfloser* 'more helpless'. The copula variant has very different types e.g. *besser* is attested after *je*. Verbless CCs have different usage (more lexicalized?)
- Extent of Use shows unsurprisingly that *je/desto+COMP* are very rare uses of the comparative, and *je COMP desto COMP* very rarely manifests itself
- Hapax productivity shows *je* and *desto* are responsible for little productivity in comparatives, but their category productivity shows they have the potential for many novel constructions, and more so for *desto* than *je*

	comp	je X	desto X	je X desto Y
Extent	1969	184	169	30
Hapax	.001378	97/565020=.000017	101/565020=.000017	√24/565020 ~ 0
Category	.006881	97/1455=.066666	101/970=.104123	√24/51=.096058