Doktorandentag
HU Berlin
13.10.2010

Novel Argument Selection: Is Lexical Semantics Enough?

Amir Zeldes

Korpuslinguistik und Morphologie, Humboldt-Universität zu Berlin

Models Explaining Novel Arguments and Productivity

Algebraic theories of grammar assume a 'words and rules' model:

- A lexically stored vocabulary (morphemes, words, multi-word units...)
- A set of rules to combine them (the grammar)

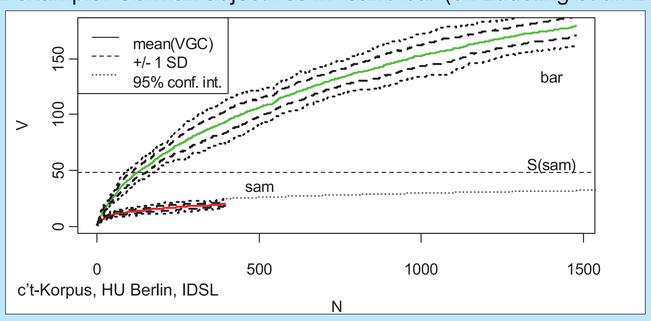
- Neologisms are explained by productively applying rules to vocabulary
- Productivity is **binary**: a pattern is productive if it results from a rule
- The class of bases (what a rule can apply to) is determined categorically, e.g. by **lexical semantics**
- The following facts are **not specified by grammar:**
  - **How often** a rule is used
  - **How many different forms** it produces
  - **How likely novel forms are**

Usage Based Models say grammar must explain **how** patterns are used

- How often, when it is selected, and what it is applied to can be **a matter of quantity or probability**
- In many models, **productivity** is seen to be **a gradient** feature, related to gradient **entrenchment** and **compositionality**

Background from Morphological Theory

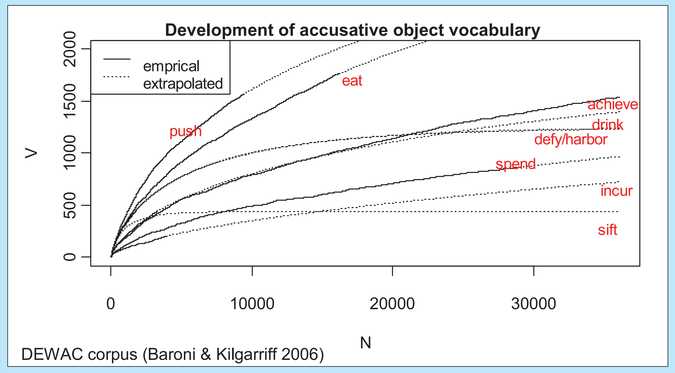
- Morphological processes can be more or less productive (Bauer 2001):
  - Neologisms in ***-tum*** are possible: *Syntaktikertum* 'syntactician-dom'…
  - but not as likely as ones in ***-keit****: Miniaturisierbarkeit* 'miniaturizability'
- Different attempts have been made to **measure productivity** (Baayen 2009)
  - Based on token frequency (**N**) - e.g. nouns in *-keit* are very frequent
  - Type frequency or vocabulary (**V**) - there are many such nouns
  - Unique, potentially novel forms (hapax legomena, **V1**) – there are many nouns found only once in large corpora
- With increasing sample size N, it becomes harder to find new types, V rises more slowly, and the probability drops that V1 increases (**P**)
- We can chart the rise of V with growing N and estimate the limit **S** of V's growth using statistical models (Evert 2004)
- A typical example: German adjectives in *-sam*/*-bar* (cf. Lüdeling et al. 2000)



Application to Syntactic Argument Selection

- Similar phenomena can be observed in syntax:
  - Verbs can be more or less **frequent (N)**
  - Some verbs have **more varied arguments** than others **(V)**
  - are more or less **likely to govern novel arguments (V1)**
- We can measure **N, V, V1, P, S** for accusative objects:

lemma	N	V	V1	P	S	$V_{N=1000}$	$P_{N=1000}$
spend	28748	862	450	0.0156	2585.051	100	0.058
sift	268	135	88	0.3283	437.7089		
push	9380	1563	796	0.0848	3023.019	398	0.276
incur	3893	203	121	0.0310	3506.464	74	0.041
harbor	1781	456	264	0.1482	1255.09	319	0.194
eat	16201	1764	917	0.0566	5377.584	323	0.201
drink	3293	444	250	0.0759	2011.245	148	0.09
defy	1705	441	260	0.1524	1245.031	307	0.191
achieve	36121	1537	759	0.0210	4343.072	190	0.117



Productivity Rankings

We get different rankings based on different criteria:

- High P means high potential productivity (novelties expected)
- High V means high realized productivity (used often so far)
- High N means high usage (forms are central to language use)
- High S means low saturation (many new uses not explored yet)

Rank	S	V	N	P	$V_{N=1000}$	$P_{N=1000}$
1	*eat*	*eat*	*achieve*	*sift*	*push*	*push*
2	*achieve*	*push*	*spend*	*defy*	*eat*	*eat*
3	*incur*	*achieve*	*eat*	*harbor*	*harbor*	*harbor*
4	*push*	*spend*	*push*	*push*	*defy*	*defy*
5	*spend*	*drink*	*incur*	*drink*	*achieve*	*achieve*
6	*drink*	*harbor*	*drink*	*eat*	*drink*	*drink*
7	*harbor*	*defy*	*harbor*	*incur*	*spend*	*spend*
8	*defy*	*incur*	*defy*	*achieve*	*incur*	*incur*
9	*sift*	*sift*	*sift*	*spend*		

Why is this important? Or: Is Lexical Semantics Enough?

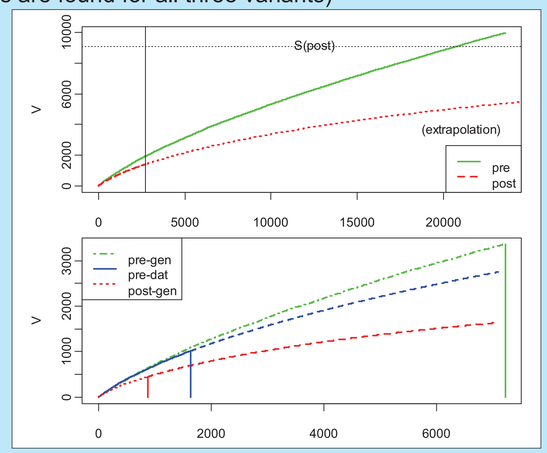
- In algebraic models **categorical distinctions** explain argument filling:
  - [+edible] <> possible object of *eat*
- In some cases, this leads to **circular argument definition**:
  - [+incurrable] <> possible object of *incur* (???)
- Do we need to **know how productive** a construction is to use it right?

The case of *wegen*

- As an example, consider German *wegen* 'because' in 3 synonymous constructions (Petig 1997, Helbig & Buscha 2001:356):
  - Preposition with genitive: *wegen des Vaters* [standard, formal]
  - Preposition with dative: *wegen dem Vater* [colloquial nonstandard]
  - Postposition with genitive: *des Vaters wegen* [formal, archaic]
- Intuitively, the **postposition is going out of use** but still productive (novel arguments are found for all three variants)

DEWAC corpus, with case ambiguities

DEWAC corpus, no case ambiguities



- prepositional forms are more productive than the postposition in all respects: not just more frequent, but **higher V, V1, P for the same N**
- Algebraic grammar cannot explain why postpositive *wegen* takes novel objects less often (semantically compatible with same objects)
- Usage based approaches assume gradient productivity, predict lower S and explain how speakers **know** not use postpositive *wegen* as readily with novel arguments
- **Hypothesis for further study**: rarity of hapax legomena leads to postposition being acquired as less productive since **speakers reproduce the input frequency distribution**

Literature:

-Bauer, L. 2001. Morphological Productivity. Cambridge: CUP.
-Baayen, R.H. 2001. Word Frequency Distributions. Dordrecht: Kluwer.
-Evert, S. 2004. A simple LNRE model for random character sequences. *Proceedings of JADT 2004*, 411-422.
-Helbig, G./Buscha, J. 2001. *Deutsche Grammatik*. Berlin: Langenscheidt.
-Lüdeling, A./Evert, S./Heid, U. 2000. On measuring morphological productivity. *Proceedings of KONVENS-2000*, 57-61.
-Petig, W.E. 1997. Genitive prepositions used with the dative in spoken German. *Unterrichtspraxis* 30, 36–39.