Amir Zeldes
Korpuslinguistik /
SFB 632 D1
Humboldt-Universität
zu Berlin

Novel Argument Realization: Semantic, Pragmatic and Conventional Productivity Effects

Linguistic Evidence 2012, Tübingen, 9-11 February 2012

Research question and intuitions

- Some verbs are more restricted than others in allowing novel or unlexicalized arguments:

  (1) *pose/represent a* ***challenge***
  (2) ?*pose/represent a* ***provocation***

- Verbs like *pose* preferably appear with collocated objects
- They occur with much fewer arguments in corpus data
- **Questions:**
  - What determines the spectrum of realized arguments?
  - Can lexical semantic classes predict argument realizations?
  - Are differences between verbs motivated by pragmatics/world knowledge?
  - Are there idiosyncratic effects that cannot be derived from verb meaning?

Semantic classes and argument selection

- **Lexical semantics** regards argument slots as realizing **semantic classes** (Katz & Fodor 1963, Jackendoff 1990), as in (3).
- The same function may be realized in constraint based grammars using specific semantic roles, e.g. **DRINKER** and **DRUNK** in **HPSG** (4):

  (3) *drink*(AGENT, PATIENT$_{[+liquid]}$)

  (4)
  $$\begin{bmatrix} \text{PHON } \langle drink \rangle \\ \text{SYNSEM} \begin{bmatrix} \text{CAT} & \begin{bmatrix} category \\ \text{HEAD} & verb \\ \text{SUBCAT} & \langle \text{NP[nom]}①, \text{NP[acc]}② \rangle \end{bmatrix} \\ \text{CONTENT} & \begin{bmatrix} \text{RELN} & drink \\ \text{DRINKER} & 1 \\ \text{DRUNK} & 2 \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

- Using such classes it is possible to account for any argument spectrum:

  (5) *pose*(AGENT, PATIENT$_{[+posable]}$)

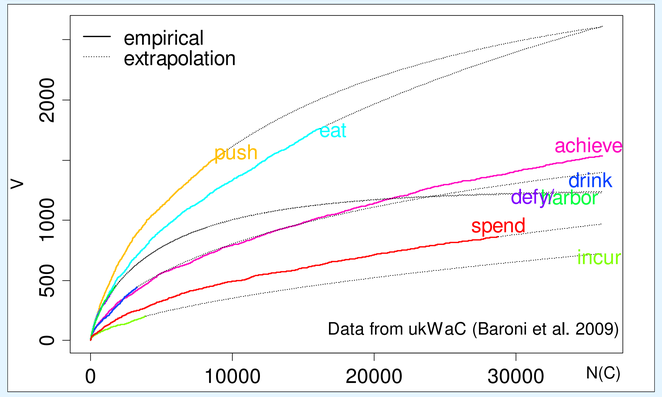
- Risk of circular logic, turning **semantic classes into a tautology** (cf. Dowty 1991)
- In order for semantic classes to predict novel arguments, **classes should be**:
  1. **Cognitively plausible**
  2. **General**, i.e. applying to as many predicates as possible
  3. **As specific as necessary**, cf. McCawley 1968: *diagonalize*(PATIENT$_{[+matrix]}$)
- Classes we define should be preserved under decomposition, cf. Jackendoff (1990):

  (6)
  $$\begin{bmatrix} \text{drink} \\ \text{V} \\ \text{— } \langle \text{NP}_j \rangle \\ [_{\text{Event}}\ \text{CAUSE}\ ([_{\text{Thing}}\ ]_i, [_{\text{Event}}\ \text{GO}\ ([_{\text{Thing}}\ \text{LIQUID}]_j, \\ \quad [_{\text{Path}}\ \text{TO}\ ([_{\text{Place}}\ \text{IN}\ ([_{\text{Thing}}\ \text{MOUTH OF}\ ([_{\text{Thing}}\ ]_i)])])])])] \end{bmatrix}$$

Methodology – productivity in argument selection

- Multiple measures are useful for comparing argument productivity (cf. Baayen 2001 on morphological productivity)
  - Frequency ***N(C)*** - Hapax legomena ***V1***
  - Vocabulary ***V*** - Estimated total vocabulary ***S*** (s. Evert 2004)
- Different rankings depending on the measure selected:

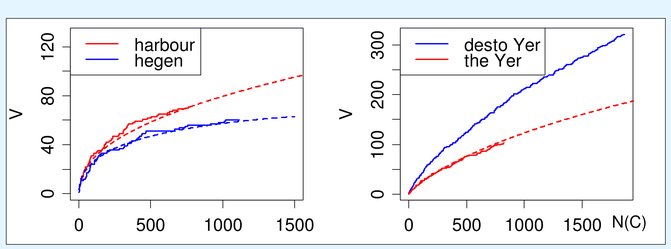
Rank	Token Frequency N(C)		Type frequency $V_{N(C)=1000}$		Hapax frequency $V1_{N(C)=1000}$		Total vocabulary S (fZM estimate)	
1	*achieve*	36121	*eat*	398	*push*	276	*eat*	5377.584
2	*spend*	28748	*push*	323	*eat*	201	*achieve*	4343.072
3	*eat*	16201	*achieve*	319	*harbour*	194	*incur*	3506.464
4	*push*	9380	*spend*	307	*defy*	191	*push*	3023.019
5	*incur*	3893	*drink*	190	*achieve*	117	*spend*	2585.051
6	*drink*	3293	*harbour*	148	*drink*	90	*drink*	2011.245
7	*harbour*	1781	*defy*	100	*spend*	58	*harbor*	1255.090
8	*defy*	1705	*incur*	74	*incur*	41	*defy*	1245.031



What does this mean? 3 case studies

1. Translational pairs

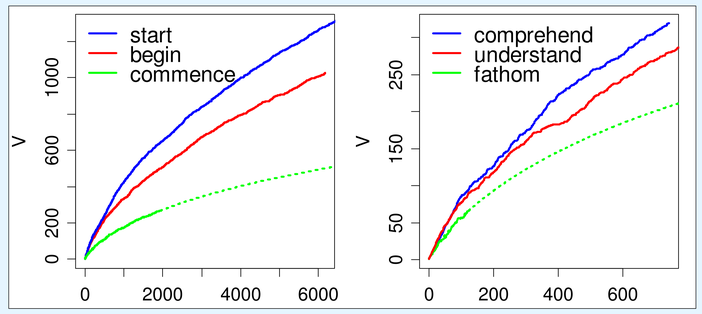
- Rankings = real world differences? (e.g. [+edible] > [+drinkable])
- If so, we expect similar productivity **cross-linguistically**
- **Counter examples** can be found in:
  - Lexeme pairs: e.g. En. *harbor* > De. *hegen,* with [+mental state]
  - Lexically unspecified constructions, e.g. De. *je Xer desto Yer* > En. *the Xer the Yer*



(Data from ukWaC and deWaC, Baroni et al. 2009)

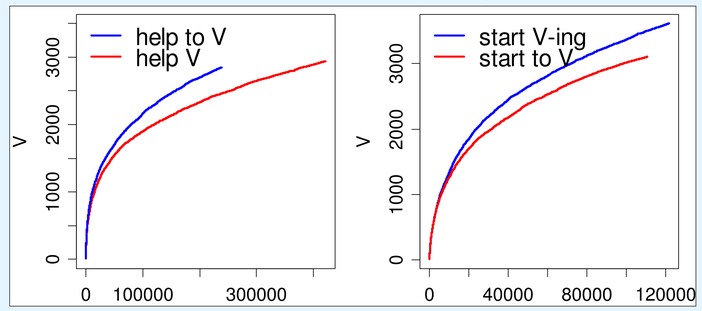
2. Near synonyms

- Synonyms are defined as **interchangeable** *salva veritate*
- Should select **same semantic class**, but again we find differences:



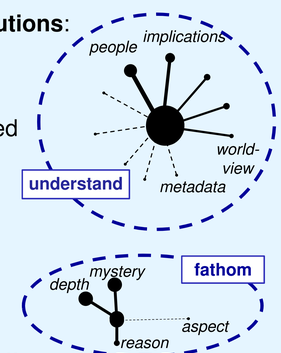
3. Syntactic alternations

- **Same head, different constructions:** again significant differences



Extensible slots in the mental lexicon – A Hebbian account

- Lexical semantics cannot explain differences in these minimal pairs
- Speakers somehow **reproduce input distributions**:
  - Argument slots with large *V*, *V1* produce more arguments in unseen data
  - I assume Hebb's Law strengthens connected representations of constructional slots and attested arguments
  - Activation by hapax legomena strengthens almost only the construction itself, without creating an entrenched argument
  - Slots with too frequent prototypes and too few rare items become identified with those arguments and tend to activate only with them



References

- Baayen, R. H. 2001. *Word Frequency Distributions.* Dordrecht: Kluwer.
- Baroni, M./Bernardini, S./Ferraresi, A./Zanchetta, E. 2009. The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora. *LRE* 43(3), 209-226.
- Dowty, D. R. 1991. Thematic Proto-Roles and Argument Selection. Language 67(3), 547-619.
- Evert, S. 2004. A Simple LNRE Model for Random Character Sequences. *Proc. of JADT 2004.* Louvain, 411-422.
- Jackendoff, R. 1990. *Semantic Structures.* Cambridge, MA: MIT Press.
- Katz, J. J./Fodor, J. A. 1963. The Structure of a Semantic Theory. *Language* 39(2), 170-210.
- McCawley, J. D. 1968. The Role of Semantics in a Grammar. In Bach, E./Harms, R. T. (eds.) *Universals in Linguistic Theory.* New York: Holt, Rinehart and Winston, 124-169.