Novel Argument Realization: Semantic, Pragmatic and Conventional Productivity Effects

Amir Zeldes
Humboldt-Universität zu Berlin
amir.zeldes@rz.hu-berlin.de

This paper aims to empirically examine the role of semantic classes in constraining the productive occurrence of novel arguments. In particular it is claimed that semantic classes and extra-linguistic knowledge about the world do not suffice in order to determine how many different arguments we can expect to see how often, and with what likelihood, in a certain syntactic slot. Corpus data is used in conjunction with statistical models from Baayen's (2001) morphological productivity paradigm and a Construction Grammar approach (CxG, see Goldberg 2006, i.a.) to show that the propensity of argument structures to admit novel lexical material is at least partly arbitrary and language-specific, and as such must be stored as part of a speaker's implicit lexico-grammatical knowledge.

It is easy to observe that some syntactic argument positions are filled by more varied lexical material than others. For example, semantic classes taken by some verbs realize a wide variety of object NP heads (e.g. *eat*([+EDIBLE]), *drink*([+LIQUID])) while others appear with a much smaller variety of objects (e.g. *incur*([+ADVERSE]), *harbour*([+MENTAL STATE])). Differences occur on multiple levels: measures based on the token frequency of each category (*N(C)* following Baayen's notation), type counts (or vocabulary size, designated by *V*), or the proportion of rare items (esp. the number of *hapax legomena*, types attested only once in large datasets, labelled *V1*) at an equal sample size will lead to different rankings, as shown in Table 1 for some verbs (cf. Bauer 2001, Baayen 2009 for a summary discussion of analogous rankings of morphological processes; numbers come from 2.25 billion words of UK English Web data, see Baroni et al. 2009).

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

Tab. 1: Rankings according to frequency, type count, and amount of hapax legomena for verbal objects, manually filtered (the latter two for equal samples of 1000 tokens).

Though the list of possible arguments for any of these verbs is neither enumerable in semantic theory nor corresponds to a closed class in reality, we could explain some contrasts pragmatically. For example, it is likely the selection of liquids we drink in daily life is more repetitive than the foods we eat, despite both classes being open and expanding. Yet for other cases, it is difficult to find a satisfying reason using only our knowledge of the world: why does one, in the language found in the corpus above, speak more often about achieving things (token frequency) but specify more different things that may be pushed (type count)? Why does *harbour* take more hapax arguments while *achieve* is much more repetitive, with more recurring arguments?

The question I will be concerned with in this paper is not whether semantic classes and world knowledge are predictive of argument diversity effects (which they certainly are), but rather whether there is empirical evidence for the hypothesis that there are some differences which cannot be reduced to semantic explanations and need to be specified for a language. To test this hypothesis, I have conducted a series of corpus studies, using the very large corpora of English and German described in Baroni et al. (2009). In each study I examine several argument structures which appear synonymous (with no formal difference in meaning, *salva veritate*) and test whether their realized argument classes differ significantly.

If we conceive of argument structure as a specification of entailments or features that a lexical item must satisfy in order to be available for realization (cf. Dowty 1991, Jackendoff 1987), then we should not expect any difference in the variety of realized arguments for such synonyms based on pragmatics alone. Decompositional approaches to lexical semantics (see Jackendoff 1990, Wierzbicka 1996, Levin & Rappaport Hovav 2005) also imply that predicates that call upon the same decomposition should take the same argument class. However, as the data in Figure 1 shows, similar argument positions behave significantly differently ($p<0.0005$), including objects of near-synonymous verbs such as English *start*, *begin* and *commence* (Panel A),[1] and even alternations using one and the same verb, such as English *start* with a gerund or *to*-infinitive complement (*start to VERB* / *start VERBing*, Panel B; see Mair 2002).

[1] Differences in register between these verbs are of course recognized, though for the purpose of determining truth values they are nearly always interchangeable. It should also be noted that higher register is not necessarily indicative of lower productivity, e.g. *understand* has more repetitive argument realization than *comprehend* in the same corpus used above. For a discussion of register and productivity see Plag et al. (1999).

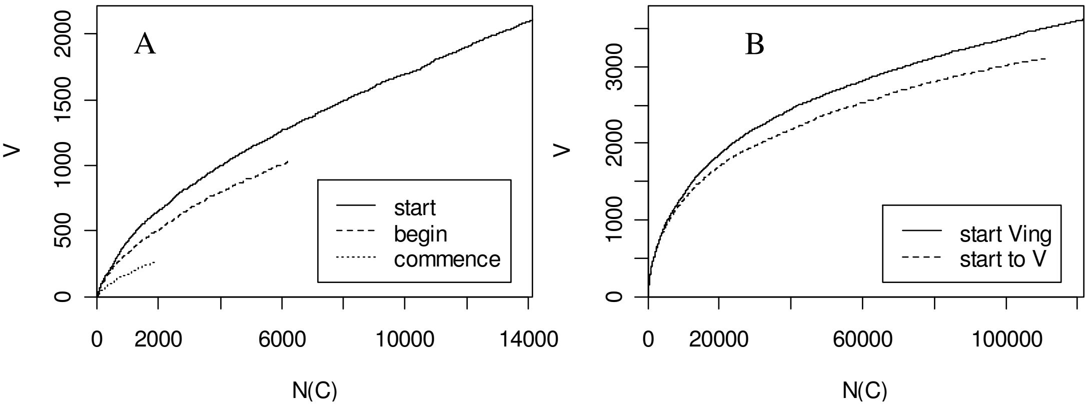

Fig. 1: Vocabulary growth curves for arguments of near-synonym verbs (A) and for synonymous argument structures of *start* with a verbal complement (B). The x-axis gives the sample size *N(C)* while the y-axis charts the amount *V* of different arguments observed thus far.

In these cases, and many more like them which will be sketched in my presentation, it is difficult to envision a formal semantic account which predicts such differences between verbs and constructions without reducing semantic classes to a circular tautology (*start* takes arguments of the class [+STARTABLE] and *begin* takes [+BEGINABLE], or worse, different classes like [+START VING-ABLE] and [+START TO V-ABLE]). But if semantic classes do not explain argument realization in usage, then what does?

Disregarding semantic classes and opting for extra-linguistic world knowledge as an explanation for these differences is also problematic, since the same extra-linguistic knowledge base should be applicable to synonymous constructions. It is also quite possible to find differences between argument extensibility in different languages for translational pairs, e.g. more varied arguments for English *harbour* with a mental state than its less flexible German counterpart *hegen* with the same meaning, and many other examples. At the same time it is understood that the way that languages divide possible arguments between verbs is arbitrary and unpredictable, so that what one *eats* in one language is *drunk* in another (e.g. soup is usually eaten in Modern Hebrew but drunk in Japanese). These findings, combined with the vocabulary studies above, highly suggest that vocabulary growth in argument realization for particular constructions is at least partly a language-specific phenomenon.

I therefore argue for lexicalized, usage-based productivity effects in argument selection which operate next to the semantic classes that determine potential argument class membership, much like the difference between the conceivable class of bases and idiosyncratic productivity ratings in morphological word formation (see Plag 1999:11-35). Different facets of productivity such as token and type frequency or the probability of encountering novel material, which are shown to be independent in my data, are all unpredictable using semantics alone, but can be estimated using statistical models developed for the study of morphological productivity (see Evert 2004; the application of such models to argument realization will be discussed in the presentation as well).

To explain the mechanism allowing differences in productivity for formally synonymous argument structures I suggest that usage information retained in the mental lexicon (or Constructicon in some CxG approaches) is stored in the form of entrenchment values for each lexicalized argument, and, additionally, entrenchment values for unlexicalized frequency bands. In other words, those hapax legomena, or other rare items, which are not stored in the lexicon or are eventually forgotten, still leave a trace on the entrenchment of hierarchically more complex constructions, contributing to their productivity. Argument slots which are attested with many infrequent items, even if these are subsequently forgotten, will be activated by each rare argument and their representation will be strengthened (cf. Bybee 1985:123), contributing to their likelihood to be selected even when an unfamiliar argument is expressed. Argument slots attested with predominantly repetitive, collocational material, will come to be identified with those arguments, leading to verbs and constructions which are avoided when novel material is to be used. The configuration of argument attestation allows us to predict productive behaviour for each slot and gives rise to fuzzy, language-specific, semantic classes that interact with self-perpetuating profiles of productive usage. In this way speakers learn a preference to realize novel, hapax arguments in constructions in which they have themselves witnessed more hapax legomena before.

References

Baayen, R. H. (2001) *Word Frequency Distributions*. (Text, Speech and Language Technologies 18.) Kluwer, Dordrecht, Boston and London.

Baayen, R. H. (2009) Corpus Linguistics in Morphology: Morphological Productivity. In: A. Lüdeling and M. Kytö, eds., *Corpus Linguistics. An International Handbook*. Vol. 2. Mouton de Gruyter, Berlin, 899-919.

Baroni, M., S. Bernardini, A. Ferraresi & E. Zanchetta (2009) The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation* 43(3): 209-226.

Bauer, L. (2001) *Morphological Productivity*. (Cambridge Studies in Linguistics 95.) Cambridge University Press, Cambridge.

Bybee, J. L. (1985) *Morphology: A Study of the Relations between Meaning and Form*. (Typological Studies in Language 9.) John Benjamins, Amsterdam and Philadelphia.

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. In: *Proceedings of JADT 2004*. Louvain-la-Neuve, Belgium, 411-422.

Goldberg, A. E. (2006) *Constructions at Work: The Nature of Generalization in Language*. Oxford University Press, Oxford.

Jackendoff, R. (1987) The Status of Thematic Relations in Linguistic Theory. *Linguistic Inquiry* 18(3): 369-411.

Jackendoff, R. (1990) *Semantic Structures*. (Current Studies in Linguistics 18.) MIT Press, Cambridge, MA.

Levin, B. & M. R. Hovav (2005) *Argument Realization*. (Research Surveys in Linguistics.) Cambridge University Press, Cambridge.

Mair, C. (2002) Three Changing Patterns of Verb Complementation in Late Modern English: A Real-time Study Based on Matching Text Corpora. *English Language and Linguistics* 6(1): 105-131.

Plag, I. (1999) *Morphological Productivity. Structural Constraints in English Derivation*. Mouton de Gruyter, Berlin and New York.

Plag, I., C. Dalton-Puffer & R. H. Baayen (1999) Productivity and Register. *English Language and Linguistics* 3: 209-228.

Wierzbicka, A. (1996) *Semantics: Primes and Universals*. Oxford University Press, Oxford.