Towards an Inventory of English Verb Argument Constructions

Matthew Brook O'Donnell
University of Michigan
500 E. Washington St.
Ann Arbor, MI 48104, USA
mbod@umich.edu

Nick Ellis
University of Michigan
500 E. Washington St.
Ann Arbor, MI 48104, USA
ncellis@umich.edu

Abstract

This paper outlines and pilots our approach towards developing an inventory of verb-argument constructions based upon English form, function, and usage. We search a tagged and dependency-parsed BNC (a 100-million word corpus of English) for Verb-Argument Constructions (VACs) including those previously identified in the pattern grammar resulting from the COBUILD project. This generates (1) a list of verb types that occupy each construction. We next tally the frequency profiles of these verbs to produce (2) a frequency ranked type-token distribution for these verbs, and we determine the degree to which this is Zipfian. Since some verbs are faithful to one construction while others are more promiscuous, we next produce (3) a contingency-weighted list reflecting their statistical association. To test whether each of these measures is a step towards increasing the learnability of VACs as categories, following principles of associative learning, we examine 20 verbs from each distribution. Here we explore whether there is an increase in the semantic cohesion of the verbs occupying each construction using semantic similarity measures. From inspection, this seems to be so. We are developing measures of this using network measures of clustering in the verb-space defined by WordNet and Roget's Thesaurus.

1 Construction grammar and Usage

Constructions are form-meaning mappings, conventionalized in the speech community, and entrenched as language knowledge in the learner's mind. They are the symbolic units of language relating the defining properties of their morphological, lexical, and syntactic form with particular semantic, pragmatic, and discourse functions (Goldberg, 2006). Construction Grammar argues that all grammatical phenomena can be understood as learned pairings of form (from morphemes, words, idioms, to partially lexically filled and fully general phrasal patterns) and their associated semantic or discourse functions: "the network of constructions captures our grammatical knowledge *in toto*, i.e. It's constructions all the way down" (Goldberg, 2006, p. 18). Such beliefs, increasingly influential in the study of child language acquisition, have turned upside down generative assumptions of innate language acquisition devices, the continuity hypothesis, and top-down, rule-governed, processing, bringing back data-driven, emergent accounts of linguistic systematicities.

Frequency, learning, and language come together in usage-based approaches which hold that we learn linguistic constructions while engaging in communication. The last 50 years of psycholinguistic research provides the evidence of usage-based acquisition in its demonstrations that language processing is exquisitely sensitive to usage frequency at all levels of language representation from phonology, through lexis and syntax, to sentence processing (Ellis, 2002). Language knowledge involves statistical knowledge, so humans learn more easily and process more fluently high frequency forms and 'regular' patterns which are exemplified by many types and which have few competitors. Psycholinguistic perspectives thus hold that language learning is the associative learning of representations that reflect the probabilities of occurrence of form-function mappings. Frequency is a key determinant of acquisition because 'rules' of language, at all levels of analysis from phonology, through syntax, to discourse, are structural regularities which emerge from learners' lifetime unconscious analysis of the distributional characteristics of the language input.

If constructions as form-function mappings are the units of language, then language acquisition involves inducing these associations from experience of language usage. Constructionist accounts of language acquisition thus involve the distributional analysis of the language stream and the parallel analysis of contingent perceptuo-motor activ-

Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics, pages 9–16,
Los Angeles, California, June 2010.

ity, with abstract constructions being learned as categories from the conspiracy of concrete exemplars of usage following statistical learning mechanisms (Bod, Hay, & Jannedy, 2003; Bybee & Hopper, 2001; Ellis, 2002) relating input and learner cognition. Psychological analyses of the learning of constructions as form-meaning pairs is informed by the literature on the associative learning of cue-outcome contingencies where the usual determinants include: (1) input frequency (type-token frequency, Zipfian distribution, recency), (2) form (salience and perception), (3) function (prototypicality of meaning, importance of form for message comprehension, redundancy), and (4) interactions between these (contingency of form-function mapping) (Ellis & Cadierno, 2009).

2 Determinants of construction learning

In natural language, Zipf's law (Zipf, 1935) describes how a handful of the highest frequency words account for the most linguistic tokens. Zipf's law states that the frequency of words decreases as a power function of their rank in the frequency table. If p_f is the proportion of words whose frequency rank in a given language sample is f, then $p_f \sim f^{-b}$, with $b \approx 1$. Zipf showed this scaling relation holds across a wide variety of language samples. Subsequent research provides support for this law as a linguistic universal: many language events (e.g., frequencies of phoneme and letter strings, of words, of grammatical constructs, of formulaic phrases, etc.) across scales of analysis follow this law (Solé, Murtra, Valverde, & Steels, 2005).

Goldberg, Casenhiser & Sethuraman (2004) demonstrated that in samples of child language acquisition, for a variety of verb-argument constructions (VACs), there is a strong tendency for one single verb to occur with very high frequency in comparison to other verbs used, a profile which closely mirrors that of the mothers' speech to these children. Goldberg et al. (2004) show that Zipf's law applies within VACs too, and they argue that this promotes acquisition: tokens of one particular verb account for the lion's share of instances of each particular argument frame; this pathbreaking verb also is the one with the prototypical meaning from which the construction is derived (see also Ninio, 1999).

Ellis and Ferreira-Junior (2009) investigate effects upon naturalistic second language acquisition of type/token distributions in three English verb-argument constructions. They show that VAC verb type/token distribution in the input is Zipfian and that learners first acquire the most frequent, prototypical and generic exemplar. (e.g. *put* in VOL [verb-object-locative], *give* in VOO [verb-object-object], etc.). Acquisition is affected by the frequency distribution of exemplars within each island of the construction, by their prototypicality, and, using a variety of psychological (Shanks, 1995) and corpus linguistic association metrics (Gries & Stefanowitsch, 2004), by their contingency of form-function mapping. This fundamental claim that Zipfian distributional properties of language usage helps to make language learnable has thus begun to be explored for these three VACs, at least. It remains an important research agenda to explore its generality across a wide range of constructions (i.e. the constructicon).

The primary motivation of construction grammar is that we must bring together linguistic form, learner cognition, and usage. An important consequence is that constructions cannot be defined purely on the basis of linguistic form, *or* semantics, *or* frequency of usage *alone*. All three factors are necessary in their operationalization and measurement. Our research aims to do this. We hope to describe the verbal grammar of English, to analyze the way VACs map form and meaning, and to provide an inventory of the verbs that exemplify constructions and their frequency. This last step is necessary because the type-token frequency distribution of their verbs determines VAC acquisition as abstract schematic constructions, and because usage frequency determines their entrenchment and processing.

This paper describes and pilots our approach. We focus on just two constructions for illustration here (V *across* n, and V Obj Obj) although our procedures are principled, generic and applicable to all VACs. We search a tagged and dependency-parsed British National Corpus (a 100-million word corpus of English) for VACs including those previously identified in the COBUILD pattern grammar project. This generates (1) a list of verb types that occupy each construction. We next tally the frequency profiles of these verbs to produce (2) a frequency ranked type-token distribution for these verbs, and we determine the degree to which this is Zipfian. Since some verbs are faithful to one construction while others are more promiscuous, we next produce (3) a contingency-weighted list which reflects their statistical association.

3 Method

As a starting point, we considered several of the major theories and datasets of construction grammar such as FrameNet (Fillmore, Johnson, & Petruck, 2003). However, because our research aims to empirically determine the semantic associations of particular linguistic forms, it is important that such forms are initially defined by bottom-up means that are semantics-free. There is no one in corpus linguistics who ‘trusts the text’ more than Sinclair (2004). Therefore we chose the Pattern Grammar (Francis et al. 1996) definition of Verb constructions that arose out of his Cobuild project.

3.1 Construction inventory: COBUILD Verb Patterns

The form-based patterns described in the COBUILD Verb Patterns volume (Francis et al. 1996) take the form of word class and lexis combinations, such as V *across* n, V *into* n and *V n n*. For each of these patterns the resource provides information as to the structural configurations and functional/meaning groups found around these patterns through detailed concordance analysis of the Bank of English corpus during the construction of the COBUILD dictionary. For instance, the following is provided for the V *across* n pattern (Francis, et al., 1996, p. 150):

> The verb is followed by a prepositional phrase which consists of across and a noun group.
>
> This pattern has one structure:
> * Verb with Adjunct.
> *I cut across the field.*

Further example sentences are provided drawn from the corpus and a list of verbs found in the pattern and that are semantically typical are given. For this pattern these are: *brush, cut, fall, flicker, flit plane, skim, sweep*. No indication is given as to how frequent each of these types are or how comprehensive the list is. Further structural (syntactical) characteristics of the pattern are sometimes provided, such as the fact that for V *across* n the prepositional phrase is an adjunct and that the verb is never passive.

For some construction patterns with a generally fixed order it may be sufficient just to specify combinations of word and part-of-speech sequences. For example, a main verb followed by *across* within 1 to 3 words (to allow for adverbial elements), followed by a noun or pronoun within a few words. To such constraints a number exceptions of what should not occur within the specified spans must be added. The variation and potential complexity of English noun phrases presents challenges for this approach. On the other hand a multi-level constituent parse tree provides more than needed. A dependency parse with word-to-word relations is well suited for the task.

3.2 Corpus: BNC XML Parsed

The analysis of verb type-token distribution in the kinds of construction patterns described in the previous section should ideally be carried out using a range of corpora in the magnitude of the tens or hundreds of millions of words as the original work is derived from the Bank of English (a growing monitor corpus of over 400 million words). These corpora should, at the least, be part-of-speech tagged to search for the pattern as specified. Further some kind of partial parsing and chunking is necessary to apply the structural constraints (see Mason & Hunston, 2004 for exploratory methodology). We chose to use the 100 million word British National Corpus (BNC) on account of its size, the breadth of genres it contains and consistent lemmatization and part-of-speech tagging. Andersen et al. (2008) parsed the XML version of the BNC using the RASP parser (Briscoe, Carroll, & Watson, 2006). RASP is a statistical feature-based parser that produces a probabilistically ordered set of parse trees for a given sentence and additionally a set of grammatical relations (GRs) that capture “those aspects of predicate-argument structure that the system is able to recover and is the most stable and grammar independent representation available” (Briscoe, et al., 2006, p. 79). The GRs are organized into a hierarchy of dependency relations, including distinctions between modifiers and arguments and within arguments between subject (sub) and complements (comp). Figure 1 shows the GRs assigned by RASP for the sentence: *The kitchen light skids across the lawn* (BNC A0U). The main verb *skids* has two arguments, a subject (ncsubj) and indirect object (iobj), and the preposition one argument (dobj).

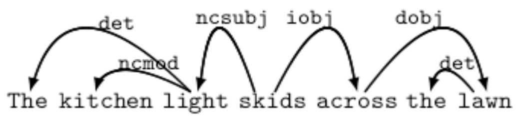

Figure 1. Example of RASP GRs

The RASP GR hierarchy does not include categories such as prepositional complement or adjunct. Figure 2 shows the GRs for another sentence containing *across* which is not an example of the V *across* n pattern. Alternate analyses might attach *across* directly to the main verb *threw*, but at least from examining BNC examples containing *across*, it appears RASP tends to favor local attachments (also for *towards* in this case).

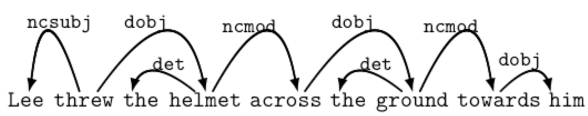

Figure 2. Example of RASP GRs

The GRs from RASP have been incorporated into the XML for each BNC sentence thereby preserving the token, part-of-speech and lemma information in the corpus.

3.3 Searching construction patterns

Our search algorithm works as follows:

1. Process each sentence in turn testing against an XPath expression to identify components in construction patterns, e.g. `.//w[@lem="across"][@pos="PREP"]/preceding-sibling::w[position()<3][@pos="VERB"][1]` finds a verb followed by *across* within 2 words.
2. Create a list of the grammatical relations where this verb functions as the head.
   i. This finds the `ncsubj` and `iobj` relations for the example sentence.
   ii. Also find GRs involving other components of pattern (e.g. *across*).
3. Check these GRs against a constraint list, e.g. make sure that
   i. only one relation where the dependent word comes after the verb (excluding verbs with both `dobj` and `iobj` or `obj2`)
   ii. the dependent of the second component matches a specific part-of-speech (e.g. *across* as head and noun as dependent).
4. For matching sentences record verb lemma.

Here we report on just two construction patterns: 1. V *across* n and 2. V n n or V Obj Obj (where n includes both nouns and pronouns). We have also run a range of similar V Prep n patterns from COBUILD, such as V *into* n, V *after* n, V *as* n. We have still to carry out a systematic precision-recall analysis, but ad hoc examination suggests that the strict constraints using the dependency relations provides a reasonable precision and the size of the corpus results in a large enough number of tokens to carry out distributional analysis (see Table 1).

Construction	Types	Tokens	TTR
V *across* n	799	4889	16.34
V Obj Obj	663	9183	7.22

Table 1. Type-Token data for V *across* n and V Obj Obj constructions

3.4 Identifying the meaning of verb types occupying the constructions

We considered several ways of analyzing the semantics the resulting verb distributions. It is important that the semantic measures we employ are defined in a way that is free of linguistic distributional information, otherwise we would be building in circularity. Therefore methods such as LSA are not applicable here. Instead, our research utilizes two distribution-free semantic databases: (1) Roget's thesaurus, a classic lexical resource of long-standing proven utility, based on Roget's guided introspections, as implemented in the Open Roget's Project (Kennedy, 2009). This provides various algorithms for measuring the semantic similarity between terms and between sentences. (2) WordNet, based upon psycholinguistic theory and in development since 1985 (Miller, 2009). WordNet classes words into a hierarchical network. At the top level, the hierarchy for verbs is organized into 15 base types (such as *move1* expressing translational movement and *move2* movement without displacement, *communicate*, etc.) which then split into over 11,500 verb synonym sets or synsets.

Verbs are linked in the hierarchy according to relations such as hypernym (to *move* is an hypernym of to *walk*), and troponym, the term used for hyponymic relations in the verb component of WordNet (to *lisp* is a troponym of to *talk*). There are various algorithms to determine the semantic similarity between synsets in WordNet which consider the distance between the conceptual categories of words, as well as considering the hierarchical structure of the WordNet (Pedersen et al. 2004).

3.5 Determining the contingency between construction form and function

Some verbs are closely tied to a particular construction (for example, *give* is highly indicative of the ditransitive construction, whereas *leave*, although it can form a ditransitive, is more often associated with other constructions such as the simple transitive or intransitive). The more reliable the contingency between a cue and an outcome, the more readily an association between them can be learned (Shanks, 1995), so constructions with more faithful verb members are more transparent and thus should be more readily acquired. Ellis and Ferreira-Junior (2009) use ΔP and collostructional analysis measures (Stefanowitsch & Gries, 2003) to show effects of form-function contingency upon L2 VAC acquisition. Others use conditional probabilities to investigate contingency effects in VAC acquisition. This is still an active area of inquiry, and more research is required before we know which statistical measures of form-function contingency are more predictive of acquisition and processing. Meanwhile, the simplest usable measure is one of faithfulness – the proportion of tokens of total verb usage as a whole that appear in this particular construction. For illustration, the faithfulness of *give* to the ditransitive is approximately 0.40; that for *leave* is 0.01.

4 Results

4.1 Evaluating the verb distribution

For the V *across* n pattern the procedure outlined in the previous section results in the following list:

come	483				
walk	203 .				
cut	199	veer	4		
run	175	whirl	4	...	
spread	146	slice	4	discharge	1
...		clamber	4	navigate	1
		.		scythe	1
				scroll	1

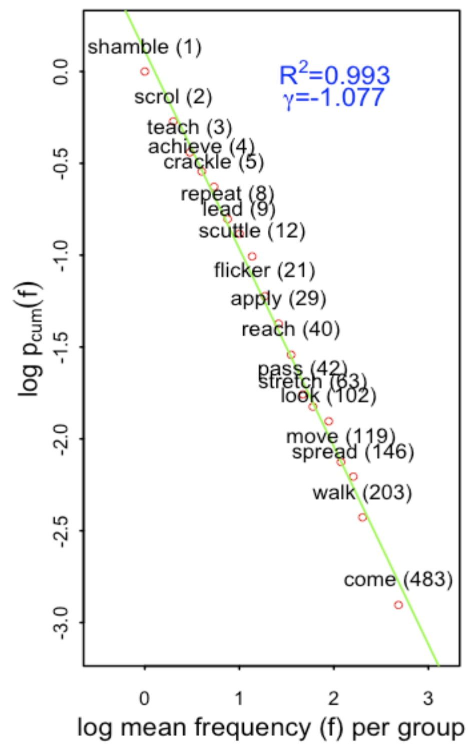

Figure 3. Verb type distribution for V *across* n

At first glance this distribution does appear to be Zipfian, exhibiting the characteristic long-tailed distribution in a plot of rank against frequency. Dorogovstev & Mendes (2003, pp. 222-223) outline the commonly used methods for measuring power-law distributions: 1. a simple log-log plot (rank/frequency), 2. log-log plot of cumulative probability against frequency and 3. the use of logarithmic binning over the distribution for a log-log plot as in 2. Linear regression can be applied to the resulting plots and goodness of fit (R^2) and the slope (γ) recorded.

Figure 3 shows such a plot for verb type frequency of the V *across* n construction pattern extracted from the parsed BNC XML corpus following the third plotting method. Verb types are grouped into 20 logarithmic bins according to their frequency (x-axis) against the logarithm of the cumulative probability of a verb occurring with or above this frequency (y-axis). Each point represents one bin and a verb from each group is randomly selected to label the point with its token frequency in parentheses. For example, the type *look* occurs 102 times in the V *across* n pattern and is placed into the 15th bin with the types *go*, *lie* and *lean*. Points towards the lower right of the plot indicate high-frequency low-type groupings and those towards the top left low-frequency high-type groupings, that is the fat- or long-tail of the distribution. Looking at the verbs given as examples of the pattern in COBUILD volume we find all but *plane* represented in our corpus search V *across* n: *brush* (12 tokens, group 9)*, cut* (199 tokens, group 18)*, fall* (57, g14)*, flicker* (21, g10)*, flit* (15, g9), *plane* (0)*, skim* (9, g8)*, sweep* (34, g12).

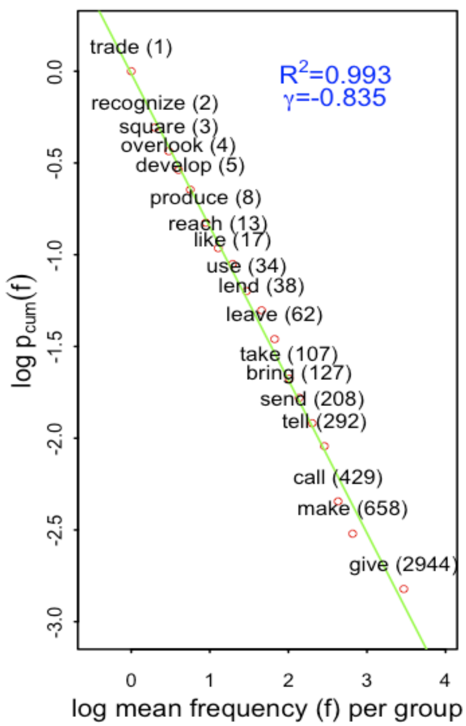

Figure 4. Verb type distribution for V Obj Obj

Figure 4 shows the plot for verb type frequency of the ditransitive V Obj Obj construction pattern extracted and binned in the same way. Both distributions can be fitted with a straight regression line (R^2=0.993). Thus we conclude that the type-token frequency distributions for these constructions are Zipfian. (In future we will investigate the other plot and fitting methods to ensure we have not smoothed the data too much through binning.) Inspection of the construction verb types, from most frequent down, also suggests that, as in prior research (Ellis & Ferreira-Junior, 2009; Goldberg et al., 2004; Ninio, 1999), the most frequent items are prototypical of the construction and more generic in their action semantics.

4.2 Evaluating the roles of frequency distribution and faithfulness in semantic cohesion

The second step in evaluating the verb distributions from the construction patterns is to compare a small set of types selected on the basis of a flat type distribution, the (Zipfian) token frequency distribution and a distribution that represents the degree to which a verb is attracted to the particular construction. First we select the top 200 types from the two VACs, ordered by token frequency. Then we sample 20 verbs from this list at random. This is the 'types list'. Next we take the top 20 types as the 'tokens list'. Finally, we calculate the tokenized faithfulness score for each type by dividing the verb's frequency in the construction by its overall frequency in the whole BNC. For example, *spread* occurs 146 times in the V *across* n pattern and 5503 times in total. So its faithfulness is 146/5503*100 = 2.65%, i.e. 1 in 38, of the instances of *spread* occur as *spread across* n. The tokenized faithfulness score for *spread* is then simply (146/5503) * 146 = 3.87, which tempers the tendency for low frequency types such as *scud*, *skitter* and *emblazon* to rise to the top of the list and is our initial attempt to combine the effects of token frequency and construction contingency. We reorder the 200 types by this figure and take the top twenty for the 'faithfulness list'. Tables 2 and 3 contain these lists for the two constructions. An intuitive reading of these lists suggests that the tokens list captures the most general and prototypical senses (*walk, move* etc. for V *across* n and *give, make, tell, offer* for V Obj Obj), while the list ordered by tokenized faith highlights some quite construction specific (and low frequency) items, such as *scud*, *flit* and *flicker* for V *across* n.

The final component is to quantify the semantic coherence or 'clumpiness' of the verbs extracted in the previous steps. For this we use WordNet and Roget's. Pedersen et al. (2004) outline six measures in their Perl WordNet::Similarity package, three (*path, lch* and *wup*) based on the path length between concepts in WordNet Synsets and three (*res*, *jcn* and *lin*) that incorporate a measure called 'information content' related to concept specificity. Tables 4 and 5 show the similarity scores that result from taking the 20 types in each of the lists in Tables 2 and 3 and generating a 20 by 20 distance matrix.

	types (sample)	**tokens**	**faithfulness**
1	scuttle	come	spread
2	ride	walk	scud
3	paddle	cut	sprawl
4	communicate	run	cut
5	rise	spread	walk
6	stare	move	come
7	drift	look	stride
8	stride	go	lean
9	face	lie	flit
10	dart	lean	stretch
11	flee	stretch	run
12	skid	fall	scatter
13	print	get	skitter
14	shout	pass	flicker
15	use	reach	slant
16	stamp	travel	scuttle
17	look	fly	stumble
18	splash	stride	sling
19	conduct	scatter	skid
20	scud	sweep	flash

Table 2. Top 20 types for V *across* n ordered by types, tokens and construction tokenized faithfulness

	types (sample)	**tokens**	**faithfulness**
1	eat	give	give
2	attend	make	call
3	feel	call	offer
4	receive	tell	make
5	miss	do	send
6	choose	offer	tell
7	affect	send	hand
8	come	show	show
9	mean	find	earn
10	provide	get	owe
11	cut	bring	cost
12	strike	ask	lend
13	prove	take	bring
14	teach	pay	do
15	refuse	allow	find
16	spare	buy	ask
17	leave	see	pay
18	wonder	hand	allow
19	permit	cost	buy
20	force	set	teach

Table 3. Top 20 types for V Obj Obj ordered by types, tokens and construction faithfulness

The figures are the mean of the values in each matrix. *Path* and *lin* values range between 0 and 1, Open Roget between 4 and 16 and the others are on varying scales where larger values indicate greater similarity. These tables show that the token distribution sample of verb types increases the semantic cohesion of the construction over a flat verbs list.

Similarity measure	Types (sampled)	Tokens (top 20)	Faithfulness (top 20)
WordNet			
path	0.163	0.387	0.245
lch	0.941	1.976	1.385
wup	0.312	0.653	0.453
res	2.473	4.673	3.748
jcn	1.033	0.383	0.190
lin	0.259	0.583	0.372
Open Roget	5.190	11.737	6.232

Table 4. Semantic similarity measures for V *across* n by types, tokens and construction faithfulness

Similarity measure	Types (sampled)	Tokens (top 20)	Faithfulness (top 20)
WordNet			
path	0.175	0.316	0.241
lch	1.008	1.654	1.299
wup	0.345	0.579	0.457
res	2.470	3.942	2.973
jcn	0.199	0.435	0.313
lin	0.308	0.558	0.406
Open Roget	7.863	13.011	10.768

Table 5. Semantic similarity measures for V Obj Obj by types, tokens and construction faithfulness

Sampling the items on the basis of their token frequency weighted for faithfulness also improves semantic homogeneity, although it does not here offer any improvement over a tokenized distribution alone. We not entirely satisfied with these measures. WordNet verb hierarchies are much flatter and bushier than those for nouns, where these measures are more successful. For verbs, distance down a synset is less telling than distance across. As a result, we are exploring other measures of the semantic similarity of verbs informed by network science. We are also exploring the use of word sense disambiguation techniques to reduce problems introduced by the rich polysemy of verbs in WordNet (e.g. *give* is assigned to 44 different synsets) and also in Roget's.

Future work

We plan to apply these methods to the full range of English VACs as described in Francis *et al* (1996) and other construction grammars too. We are particularly interested in whether the inventory represents an optimal partitioning of verb semantics, starting with basic categories of action semantics and proceeding to greater specificity via Zipfian mapping. We are also interested in extending these approaches to learner language to investigate whether first and second language learners' acquisition follows the construction distributional profiles and whether the factors outlined in Goldberg et al. (2004) facilitate acquisition.

There have been suggestions that Zipfian type-token frequency distributions are essentially uninteresting artifacts. For each motivated construction identified along the lines described in 3.3, we have begun to make matching random control distributions generated as a random selection of verb types of comparable *n* types and tokens (yoked ersatz-controls). For each of our outcome measures, we will compare the various scores for VAC verb-types gathered on the principled basis of construction-grammar against those for their controls.

Conclusions

Meanwhile, these pilot studies show some promise in these methods towards an English verb grammar operationalized as an inventory of VACs, their verb membership and their type-token frequency distributions, their contingency of mapping, and their semantic motivations.

References

Andersen, Ø. E., Nioche, J., Briscoe, T., & Carroll, J. (2008). The BNC Parsed with RASP4UIMA. *Proceedings of the Sixth International Language Resources and Evaluation (LREC08)*, 28-30.

Bod, R., Hay, J., & Jannedy, S. (eds.). (2003). *Probabilistic linguistics*. Cambridge,MA: MIT Press.

Briscoe, E., Carroll, J., & Watson, R. (2006). *The Second Release of the RASP System.* Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions, Sydney, Australia.

Bybee, J., & Hopper, P. (Eds.). (2001). *Frequency and the emergence of linguistic structure*. Amsterdam: Benjamins.

Dorogovstev, S.N., & Mendes, J.F.F. (2003). *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford: Oxford University Press.

Ellis, N.C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition, 24* (2), 143-188.

Ellis, N.C., & Cadierno, T. (2009). Constructing a second language. *Annual Review of Cognitive Linguistics, 7* (Special section).

Ellis, N.C., & Ferreira-Junior, F. (2009). Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics*, 111-139.

Fillmore, C. J., Johnson, C. R., & Petruck, M. R. L. (2003). Background to Framenet. *International Journal of Lexicography, 16*, 235-250.

Francis, G., Hunston, S., & Manning, E. (Eds.). (1996). *Grammar Patterns 1: Verbs. The COBUILD Series*. London: Harper Collins.

Goldberg, A.E. (2006). *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.

Goldberg, A.E., Casenhiser, D.M., & Sethuraman, N. (2004). Learning argument structure generalizations. *Cognitive Linguistics,* 15, 289–316.

Gries, S.T., & Stefanowitsch, A. (2004). Extending collostructional analysis: a corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics, 9*, 97-129.

Kennedy, A. (2009). The Open Roget's Project: Electronic lexical knowledge base. Retr. 1st March, 2010: http://rogets.site.uottawa.ca/index.shtml

Mason, O., & Hunston, S. (2004). The automatic recognition of verb patterns: A feasibility study. *International Journal of Corpus Linguistics, 9*, 253-270.

Miller, G. A. (2009). WordNet - About us. Retrieved March 1, 2010, from http://wordnet.princeton.edu

Ninio, A. (1999). Pathbreaking verbs in syntactic development and the question of prototypical transitivity. *Journal of Child Language, 26*, 619-653.

Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004). *WordNet::Similarity – Measuring the Relatedness of Concepts.* Proceedings of Fifth Annual Meeting of the North American Chapter of the Association of Computational Linguistics (NAACL 2004).

Shanks, D. R. (1995). *The psychology of associative learning*. New York: Cambridge University Press.

Sinclair, J. (2004). *Trust the text: Language, corpus and discourse*. London: Routledge.

Solé, R. V., Murtra, B., Valverde, S., & Steels, L. (2005). Language Networks: their structure, function and evolution. *Trends in Cognitive Sciences, 12*.

Stefanowitsch, A., & Gries, S.T. (2003). Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics, 8*, 209-243.

Zipf, G. K. (1935). *The psycho-biology of language: An introduction to dynamic philology*. Cambridge, MA: The M.I.T. Press.