John Benjamins Publishing Company

This is a contribution from *Corpora, Grammar and Discourse. In honour of Susan Hunston*.
Edited by Nicholas Groom, Maggie Charles and Suganthi John.
© 2015. John Benjamins Publishing Company

CHAPTER 2

Using COBUILD grammar patterns for a large-scale analysis of verb-argument constructions

Exploring corpus data and speaker knowledge

Ute Römer[1], Matthew B. O'Donnell[2] and Nick C. Ellis[3]
[1]Georgia State University / [2]University of Pennsylvania / [3]University of Michigan

This paper takes patterns identified in *COBUILD Grammar Patterns 1: Verbs* (Francis et al. 1996) as a starting point for the systematic, large-scale analysis of English verb-argument constructions (VACs), using both corpus/computational methods and psycholinguistic experiments. We work in an iterative cycle to define, search, review and refine patterns to retrieve VACs from a parsed version of the BNC and examine the distributions of the verb types and their token frequencies for each VAC. The findings allow us to make predictions regarding language users' knowledge of verbs in constructions. We then test these predictions in psycholinguistic experiments, in which native and non-native speakers of English think of the first word that comes to mind to fill the V slot in a particular VAC frame. We compare the results from the experiments and the corpus analysis in terms of verb selection preferences. This research demonstrates the productive synergy of corpus linguistic and psycholinguistic methods and findings.

1. Introduction: Analysing verb-argument constructions (VACs) at scale

Corpus linguistics has shown that language is highly patterned. Written sentences and spoken utterances are made up to a large extent of fixed or semi-fixed elements variously referred to as clusters, phrases, phraseological items, chunks, lexical bundles, n-grams, collocational frameworks, formulaic sequences, multi-word units, or constructions. Construction Grammar suggests a fixed form-meaning correspondence and argues that combinations of words (constructions) carry meanings as a whole (Goldberg 2003, 2006). An oft-cited example of this is the

DOI 10.1075/scl.73.03rom

'verb object object' (V obj obj) or ditransitive construction, in which a verb form is followed by an indirect and a direct object, as in 'we gave Susan a book for her birthday'. Part of our linguistic knowledge is the knowledge of what kinds of verbs may or may not occur in this construction. We know which lexical items the ditransitive construction tends to select. Also part of our linguistic knowledge is the knowledge of a verb's preferred complementation patterns or subcategorisation frames. We know which constructions the verb GIVE tends to occur in. Even when we are faced with a novel utterance that contains a nonsense verb such as 'they spugged her a present', we are able to identify SPUG as a verb that expresses some kind of transfer. SPUG here inherits its interpretation of meaning from echoes of the verbs we usually encounter in this construction: GIVE, MAKE, TELL, TAKE, and SEND. We assume that, because it occurs in the ditransitive, SPUG is semantically related to these verbs.

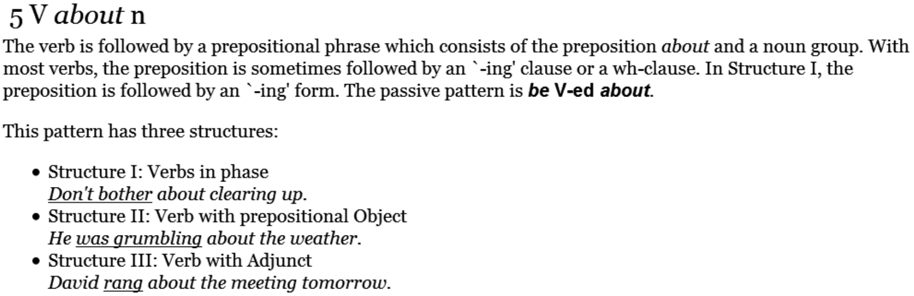
Small sets of patterns including 'V obj obj' have been studied in Construction Grammar and in first and second language acquisition (e.g. Ellis & Ferreira-Junior 2009; Goldberg 2006; Goldberg, Casenhiser & Sethuraman 2004). These studies concluded that, based on samples of native speaker and learner data, there is a strong tendency for one single verb to occur with a particularly high frequency in comparison to other verbs, and that the overall distribution of verbs in patterns or constructions follows Zipf's law, which states that the frequency of words decreases as a power function of their ranks in the frequency table (Zipf 1935). The studies show how the frequencies of verbs influence acquisition, and how Zipfian distributional properties of language usage help make language learnable, both for first and second language learners. The findings are revealing but have yet to be backed up by evidence from more constructions and larger datasets.

In a collaborative project among psycho-, corpus, and computational linguists, we are investigating the use and acquisition of a large number of verb-argument constructions (VACs) at scale, through corpus analyses and psycholinguistic experiments. Our aims are to empirically determine the type and token frequencies of verbs in constructions, their semantic associations, and the entrenchment of VACs in the native speaker's and the second language learner's mind. We have taken a large sample of patterns identified and discussed in *COBUILD Grammar Patterns 1: Verbs* (Francis et al. 1996) as a starting point for a systematic analysis of VACs in the 100-million word British National Corpus (BNC). In this paper, we first discuss the method we developed to define, review and refine search strings that allow us to retrieve VACs from a parsed version of the BNC with a high degree of accuracy in terms of precision and recall (Section 2). We then present some initial results from the corpus analysis (Section 3) and compare the corpus findings with observations made in psycholinguistic experiments on native speaker and learner associations of verbs and constructions (Section 4). We close with a summary of our findings and thoughts on future research on VAC usage and acquisition.

2. From COBUILD patterns to corpus VACs

As previously mentioned, our research aims to empirically determine not just the type and token frequencies, but also the semantic associations of verbs and constructions. It is therefore important to initially define the forms that will be analysed in a semantics-free, bottom-up manner. We chose the definition of VACs presented in Volume 1 of the Grammar Patterns series (Francis et al. 1996) that came out of the COBUILD project in corpus- and pedagogy-driven lexical computing – a project in which the corpus came first and the text was to be trusted (Sinclair 2004).

Francis et al. (1996) contains over 700 patterns of varying complexity, organised by verb type and part of speech of the pattern constituents. For example, Chapter 2 deals with "Simple Patterns with Prepositions and Adverbs" (e.g. **V *about* n**), while Chapter 9 covers "Verb Patterns with it" (e.g. **it V *that***). For each of the patterns, the book provides information on their structural configurations and the meaning groups formed by the verbs that occur in the patterns in the 300 million word version of the Bank of English corpus. The first part of the entry for the **V *about* n** pattern is presented in Figure 1. For each of the structural realisations of the pattern, the book provides corpus-derived example sentences and information on verb preferences. In the case of **V *about* n**, only two verbs (BOTHER and FORGET) are listed for Structure I. Verbs found in Structure II patterns belong to three meaning groups: the 'TALK' group, the 'THINK' group and the 'LEARN' group. Structure III verbs (including PHONE and WRITE IN) are concerned with telephone or written communication (Francis et al. 1996: 146–150). As Hunston and Francis (1999: 36) note, the descriptions in the Grammar Patterns volumes aim at a comprehensive coverage of all the lexical items that occur in a particular pattern (based on the Bank of English). For some patterns that have especially long lists of verbs, however, only the most frequent verbs are included. The entries do not indicate *how* frequent each of the listed verb types are.



5 V *about* n

The verb is followed by a prepositional phrase which consists of the preposition *about* and a noun group. With most verbs, the preposition is sometimes followed by an `-ing' clause or a wh-clause. In Structure I, the preposition is followed by an `-ing' form. The passive pattern is ***be* V-ed *about***.

This pattern has three structures:

- Structure I: Verbs in phase
  Don't bother about clearing up.
- Structure II: Verb with prepositional Object
  He was grumbling about the weather.
- Structure III: Verb with Adjunct
  David rang about the meeting tomorrow.

Figure 1. Part of the *COBUILD Grammar Patterns* entry for **V *about* n** (Francis et al. 1996: 145)

2.1 Defining search graphs from COBUILD descriptions

The definitions of VACs in the COBUILD descriptions, such as the one in Figure 1, make use of grammatical/syntactical categories such as prepositional phrase, noun group and *wh*-clause. Such categories involve combinations of words and are, to one degree or another, abstractions from the lexical level of text. They are, therefore, not easy to capture through a simple concordance search with any degree of precision.[1] Because of this, we decided to make use of a parsed corpus for the VAC searches. We selected a dependency based analysis as it does not impose constituents over words in a sentence but does capture the primary word-to-word relations necessary for identifying VACs. Andersen et al. (2008) ran the whole BNC-XML through an NLP pipeline including the RASP parser (Briscoe et al. 2006) to produce a dependency parsed version of the corpus, with a separate level of XML annotation indicating dependency pairs. To make searching across word, lemma, part-of-speech and dependency levels easier, we transformed this XML into GraphML (an XML representation of a property graph, with words as nodes and edges as dependency relations). Figure 2 shows a visual representation of the annotation of the sentence *I know he fantasises about it* (BNC G2V.1676). The numbers indicate word position, followed by lexical form. Below this are the POS tags, first the simplified word class, then the CLAWS tag separated by a dash from the RASP POS tag. The arrows indicate dependency relations, pointing from head word to dependent word (with the arrow head), e.g. *know* is the head of *I* through an 'ncsubj' (subject) grammatical relation (see Briscoe et al. 2006).

The property graph representation allows for searches to be defined for words (nodes) on any of their properties (lexical and lemma forms and/or POS categories) and for required and disallowed relations between these words. For example, for **V about n**, we would begin with the node in the graph that represents *about* that also has a direct object grammatical relation to the head of a noun phrase. The head could be a noun or a pronoun. Also the *about* node should be the dependent of a verbal node. This is again best illustrated visually. Figure 3 shows the initial search graph corresponding to this process for the **V *about* n** pattern.

Node 'n1' is the search start point and has a restriction on the lemma property to equal *about*. From here, dependent grammatical relations from *about* are queried, looking for one with a value 'dobj' (direct object). The dependent node in this relation ('n3') needs to be a noun ('N') or pronoun ('PN') (the 'c5' property is the CLAWS POS tag for a word) and is checked through a regular expression match. If the search reaches this point, then half of the match requirements have been met: *about* is the head of a prepositional phrase with a noun group

[1] Many of these issues and some tentative solutions for automatically retrieving verb patterns are discussed in some detail by Mason and Hunston (2004).

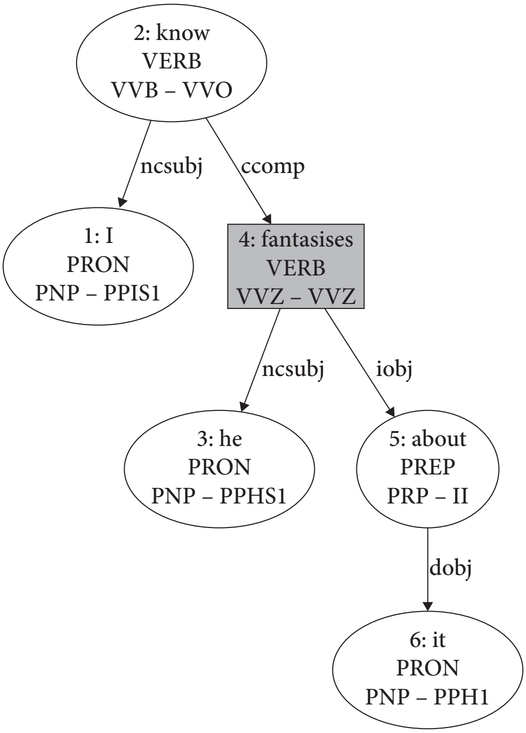


Figure 2. Representation of RASP dependency annotation of BNC G2V.1676

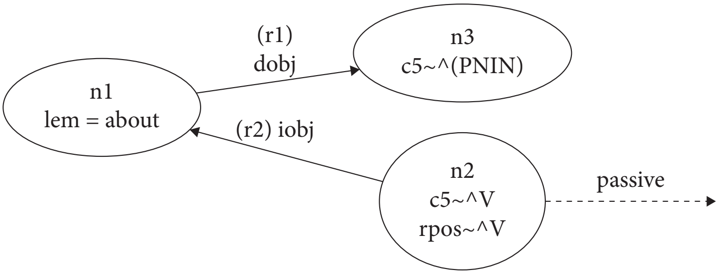


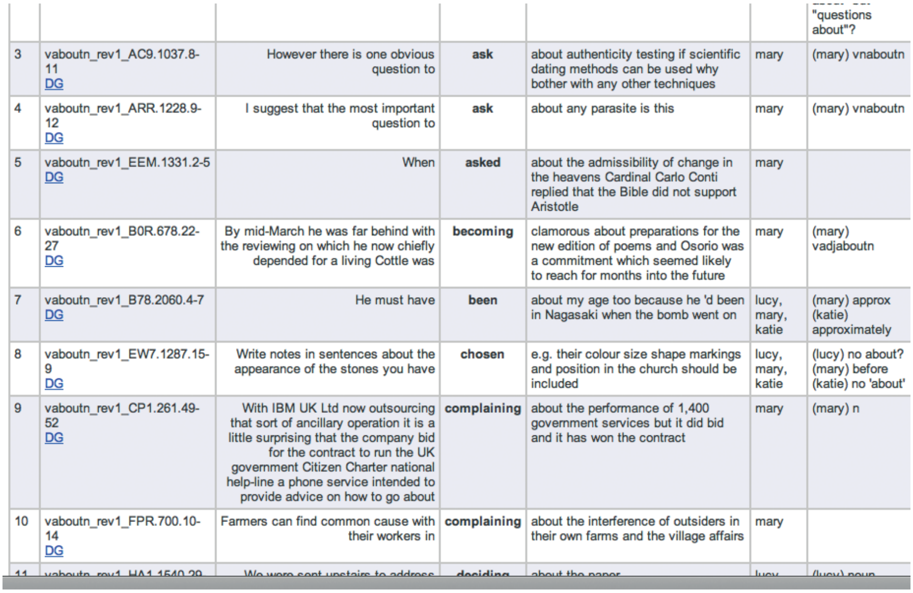
Figure 3. Representation of first stage search graph for the **V *about* n** VAC

object. Next the search looks for incoming edges (dependent relations) to *about* (n1) and specifically one with value 'iobj' (indirect object), since Francis et al. (1996: 145) state "the verb is followed by a prepositional phrase", pointing out that this may be either a prepositional object (Structure II) or an adjunct (Structure III). If such an edge exists, the node 'n2' (head word of the 'iobj' GR) is checked to see if both POS taggers agree it is a verb. The final check is to ensure that this verb is not in a passive construction (the dashed edge indicates a disallowed relation) as this is specifically excluded in the description of **V *about* n** in Francis et al. (1996).

We found that using a visual specification for a VAC search pattern provides a useful communicative bridge between the computational and linguistic members of our team. Once the search is drawn in this way, it is transferred into a descriptive XML markup that defines how the search should be carried out (technically defining a graph traversal – a record of 'walking' through the dependency structure for the sentence from word to word along the grammatical relation paths – that will result in a successful match between the VAC pattern and a sentence in the corpus). This XML is used by a Python script to search the GraphML database and returns hits of sentences matching the pattern.

2.2 Checking precision and recall of VAC searches

It is inevitable that real language in the corpus will turn out to be more complex and varied than anticipated in the specification of an abstract structure search. So the next step in our process was to carry out a precision analysis on a random sample of 500 sentences retrieved by the search. We developed a web-based interface that allowed our precision checkers to independently view these 500 sentences and compare them with the COBUILD description of the respective VAC (see Figure 4).



						"questions about"?
3	vaboutn_rev1_AC9.1037.8-11 DG	However there is one obvious question to	**ask**	about authenticity testing if scientific dating methods can be used why bother with any other techniques	mary	(mary) vnaboutn
4	vaboutn_rev1_ARR.1228.9-12 DG	I suggest that the most important question to	**ask**	about any parasite is this	mary	(mary) vnaboutn
5	vaboutn_rev1_EEM.1331.2-5 DG	When	**asked**	about the admissibility of change in the heavens Cardinal Carlo Conti replied that the Bible did not support Aristotle	mary	
6	vaboutn_rev1_B0R.678.22-27 DG	By mid-March he was far behind with the reviewing on which he now chiefly depended for a living Cottle was	**becoming**	clamorous about preparations for the new edition of poems and Osorio was a commitment which seemed likely to reach for months into the future	mary	(mary) vadjaboutn
7	vaboutn_rev1_B78.2060.4-7 DG	He must have	**been**	about my age too because he 'd been in Nagasaki when the bomb went on	lucy, mary, katie	(mary) approx (katie) approximately
8	vaboutn_rev1_EW7.1287.15-9 DG	Write notes in sentences about the appearance of the stones you have	**chosen**	e.g. their colour size shape markings and position in the church should be included	lucy, mary, katie	(lucy) no about? (mary) before (katie) no 'about'
9	vaboutn_rev1_CP1.261.49-52 DG	With IBM UK Ltd now outsourcing that sort of ancillary operation it is a little surprising that the company bid for the contract to run the UK government Citizen Charter national help-line a phone service intended to provide advice on how to go about	**complaining**	about the performance of 1,400 government services but it did bid and it has won the contract	mary	(mary) n
10	vaboutn_rev1_FPR.700.10-14 DG	Farmers can find common cause with their workers in	**complaining**	about the interference of outsiders in their own farms and the village affairs	mary	
11	vaboutn_rev1_HA1.1540.29	We were sent upstairs to address	**deciding**	about the paper	lucy	(lucy) noun

Figure 4. Precision analysis interface shown reviewing a sample of **V *about* n** sentences

The precision checkers then marked errors and were able to leave specific comments on each BNC example. In Figure 4 you can see that annotator Mary has marked sample sentence 6, *Cottle was becoming clamorous about preparations for*, as an error and suggested that it is actually an instance of the **V adj *about* n** VAC. There is a link by each sentence allowing the dependency graph for the sentence to be displayed (see Figure 5).

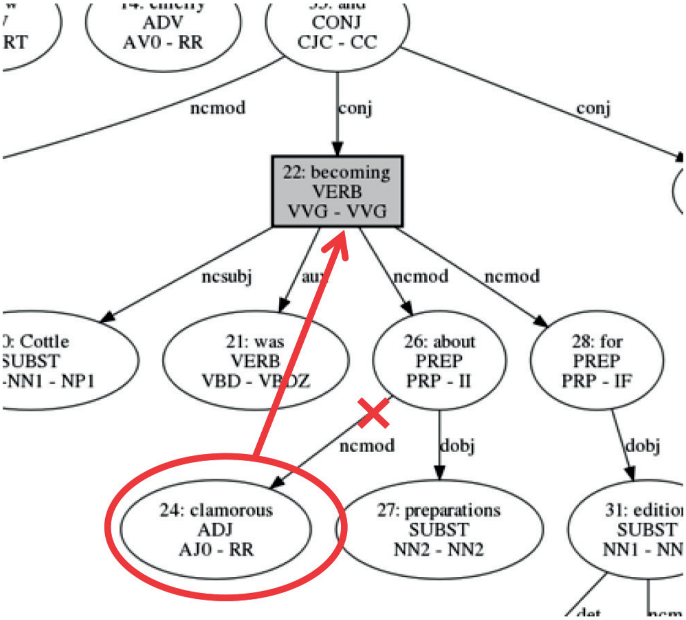


Figure 5. Representation of a sentence incorrectly identified as an instance of the **V *about* n** VAC

Examining this dependency graph reveals that in this instance the error results from the way in which RASP has chosen to classify the adjective *clamorous* as a 'ncmod' (non-clausal modifier) of the preposition *about* instead of attaching it to the verb form *becoming*. If such attachment errors resulting from tendencies (or quirks) of the parser appear consistently, they can be controlled for by adjusting the search to exclude them. The number of errors is used to calculate a precision figure. For example, for the initial search graph for **V *about* n** the annotators marked 104 out of the 500 sample sentences as errors, which gives a precision score of 79.2%.

The precision analysis provides an indication of how well the search graph specification for a particular VAC matches the definition from COBUILD. The other component of analysing accuracy – recall analysis – determines what proportion of the actual instances of a VAC in the corpus are identified by the search graph definition. Because of the variability and complexity of real language in a corpus, the abstract specification of a VAC pattern in terms of grammatical dependency relations is likely to miss some unanticipated but common patterns. We carried out recall analyses using a very simple search, i.e. just the single word *about* for the **V *about* n** VAC, in the BNCweb interface (Hoffmann et al. 2008). Each of two annotators carried out this search and each reviewed 500 separate results, which were randomly selected. Using the categorisation feature for KWIC lines in BNCweb, correct instances of the VAC were marked and these were exported and compared with the results of the BNC VAC search.

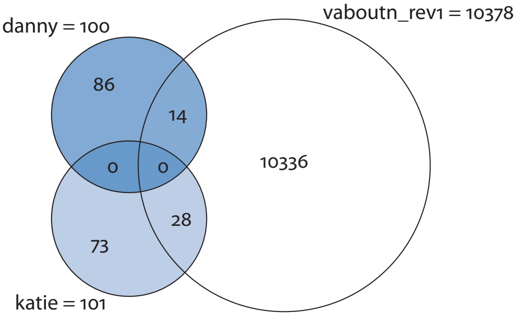


Figure 6. Recall results for initial search of **V *about* n**

Figure 6 illustrates this comparison. For the **V *about* n** VAC, two coders, Danny and Katie, carried out recall analysis of 500 randomly selected sentences containing *about*. Danny identified 100 of his sentences as genuine instances of the VAC; Katie found 101 of her sentences to be genuine **V *about* n** hits. They had no sentences in common. The corpus search based on our initial search graph for **V *about* n** retrieved 10,378 sentences. Forty-two of the 201 sentences in the recall set were among these results, giving a recall figure of 20.9%. The recall figure in particular is somewhat disappointing and should be improved in order to be able to state broad coverage in our claims regarding verbs in VACs. However, we chose to prioritise improving precision (rather than recall) as much as possible so that the statistical and semantic interpretation of our results (see Section 3) would be as reliable as possible for the subset of retrieved corpus instances. The combined F-score (using α=0.5) is therefore 1/(0.5/0.792 + 0.5/0.209) = 0.331.[2]

2.3 Refining the search graphs

The web interfaces used for both the precision and recall analyses provide links to the dependency graphs for all sentences analysed. This makes it possible to easily

[2] The F-score is a measure of the accuracy of a test or procedure which considers both its precision and recall results.

inspect how a sentence was analysed by the RASP parser and the POS taggers in order to discover why an incorrect sentence was included in our search results (see Figure 5) or why a correct one was excluded.

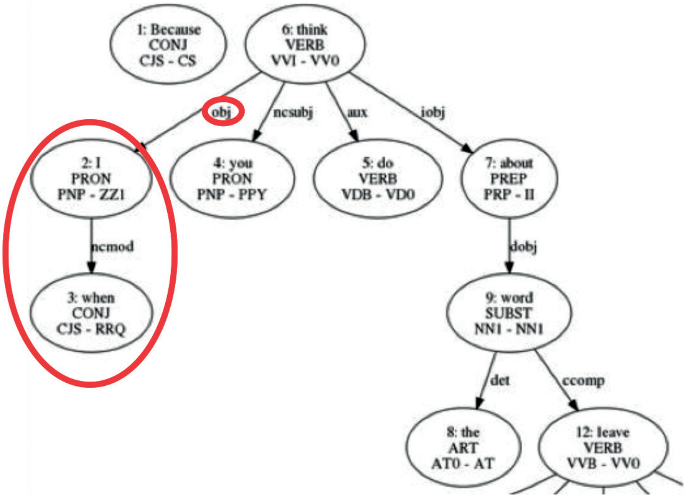


Figure 7. Representation of a true **V *about* n** sentence excluded by the initial search graph

For instance, the recall analysis of **V *about* n** identified the sentence, *Because I when you do **think about the word** then you leave some of the letters out* (BNC FMF.494), which the original search excluded. This is an instance of speech with a false start and switch of person reference from *I* to *you*. RASP clearly has difficulty assigning grammatical relations in cases like this. RASP seems to have falsely assigned an obj relation between *think* and *I* here (see Figure 7) and thus discounted this sentence as an instance of **V *about* n**, classifying it instead as **V n *about* n**.

In the review process, we collected notes on potential changes to the search graph that would exclude consistent errors found in the precision analysis and others that would broaden the coverage of the search discovered during recall analysis. Many of these potential changes stemmed from noticing the ways in which the RASP parser dealt with certain words and grammatical relations. For instance, we initially assumed that the relation between the preposition (e.g. *about*) and the head of the noun phrase would always be a 'dobj' (direct object) relation. However, through review of the sentences found in the recall analysis but not captured in the search, we discovered that complement grammatical relations, 'ccomp' and 'xcomp', could also occur. Similarly, we found that conjunctions creating complex noun (1) and verb phrases (2) result in structures where the conjunctions become head words intervening between the preposition and noun or verb phrase heads (see Figure 8).

(1) …people who (VP write (PP about (NP psychoanalyses and the social sciences NP) PP) VP)… (BNC HUK.218)

(2) He's fucking mad, Simon, he always makes people (VP (VP think and talk VP) (PP about (NP these mad things NP) PP) VP)… (BNC KC7.516)

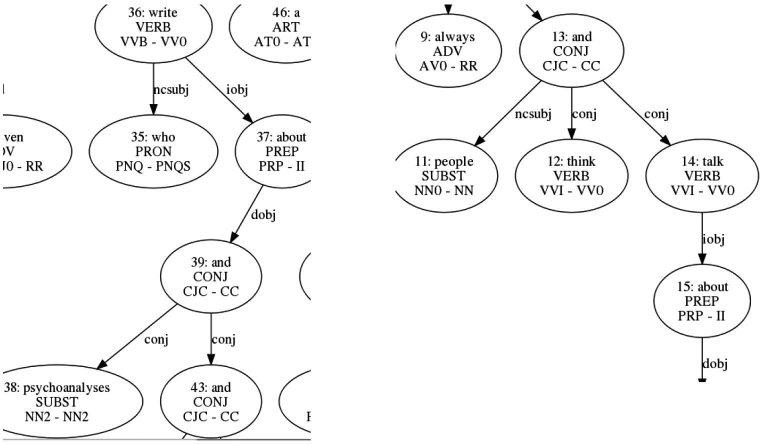


Figure 8. Search complexities introduced by complex conjunctive relations in noun and verb phrases

In addition, we found further grammatical relations and lexical combinations that needed to be excluded to ensure retrieved instances were actual instances of the VAC. For example, in the **V *about* n** VAC the verb should not have any additional object grammatical relations (dobj, obj, iobj) aside from the one between the verb and *about*. Lexical phrases *sure about*, *just about*, *round* or *around about* are not recognised as units by the parser but should be excluded from consideration.

Figure 9 shows the search graph for **V *about* n** after 3 cycles of precision, recall and revision (compare to Figure 3). You can see how the conjunctive relations discussed above have been allowed for between the verb ('n2') and *about* ('n1') and between *about* and the head of the noun phrase ('n3' or 'n5'). Notice how the part-of-speech regular expressions have expanded for this node to allow for determiners (*about this*) and -ing forms (*about clearing up*), as in Structure I of the description in Francis et al. (1996). The dashed edges and attached nodes indicate relations and patterns that are disallowed and should cause the search to make no match.

2.4 Balancing precision and recall against fidelity to COBUILD definitions

Given our primary aim of investigating patterns of verbal usage in VACs from a quantitative perspective, we opted to prioritise precision over recall. This is the

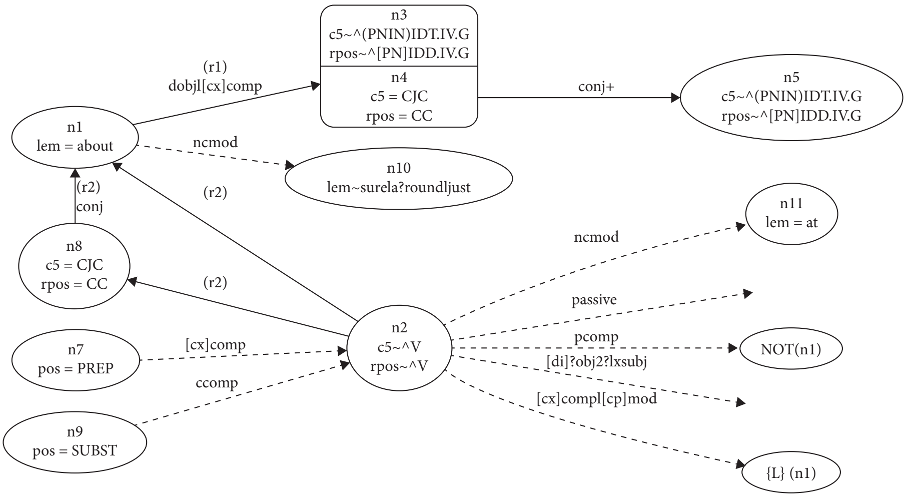


Figure 9. Representation of revised search graph for the **V** ***about*** **n** VAC

reverse of the usual practice adopted in corpus linguistic research (e.g. Hoffmann et al. 2008), which advocates an approach that produces wide coverage (i.e. high recall) and the use of manual correction, such as would be the case in a KWIC-centered analysis (e.g. Sinclair 2003). The reasoning behind this choice was that, given the size and linguistic diversity of the BNC as well as the restricted nature of our search definitions, we expect the results to be highly representative of specific VACs defined according to surface and syntactic criteria. Further, in order to carry out the kinds of analysis detailed in Section 3, including measuring the contingency of specific verb types to specific VACs and the semantic coherence of verb types within constructions, precise results of a reasonable size are more useful than a larger set of noisy results. Table 1 shows summary data, including precision and recall statistics, for 18 VACs drawn from the second chapter of Francis et al. (1996) – all are of the form 'V prep n'. For these 18 VACs after three cycles of search – precision-recall – refine, we achieved a mean precision of 78%, and recall of 53%, giving a combined F-score of 0.612. (The best value for an F-score is 1 and the worst score is 0.)

The use of a parsed corpus and syntactic search constraints, as detailed in the previous sections, does mean that our VACs, although drawn from the COBUILD patterns (Francis et al. 1996), are in many cases restricted versions of those patterns. In the same way that definitions of constructions in Construction Grammar (such as Goldberg 2003, 2006) include semantic constraints (e.g. an oblique encoding direction, means or manner), we often found semantic elements in the COBUILD definitions, particularly those distinguishing the different structures for a pattern, that we were unable to capture. The COBUILD definitions also contain a considerable richness and level of detail resulting from the close inspection of KWIC lines on which they are based that we have not tried to reproduce in our analysis. We think, however, that our VAC definitions and the resulting datasets are adequate (in terms of quality and quantity) for the purposes of the present project.

3. Initial results: VACs in a corpus

Table 1 shows the 18 selected constructions, the number of verb types that occupy them, the total number of tokens found, type-token ratios, and precision and recall figures.

Table 1. Type-token data for 18 'V prep n' VACs drawn from Francis et al. (1996) and retrieved from the BNC

VAC	Types	Tokens	TTR	Lead verb type (-BE)	Precision	Recall	F-score
V *about* n	908	24244	3.75	TALK	0.778	0.454	0.573
V *across* n	669	5261	12.72	COME	0.916	0.614	0.735
V *against* n	838	8978	9.33	LEAN	0.888	0.600	0.716
V *among* pl-n	478	2859	16.72	LIVE	0.816	0.681	0.743
V *around* n	799	5243	15.24	LOOK	0.806	0.515	0.628
V *as* n	1431	22857	6.26	ACT	0.482	0.516	0.498
V *between* pl-n	852	8300	10.27	DISTINGUISH	0.850	0.594	0.699
V *for* n	2281	90980	2.51	LOOK	0.798	0.504	0.618
V *in* n	3573	190370	1.88	LIVE	0.756	0.333	0.463
V *into* n	1689	50070	3.37	GO	0.892	0.621	0.733
V *like* n	1232	15985	7.71	LOOK	0.742	0.339	0.465
V *off* n	295	1603	18.40	GO	0.764	0.031	0.060
V *of* n	1090	44418	2.45	THINK	0.596	0.527	0.559
V *over* n	1516	19710	7.69	TAKE	0.612	0.574	0.592
V *through* n	1418	21583	6.57	GO	0.892	0.703	0.786
V *towards* n	639	8005	7.98	MOVE	0.928	0.720	0.811
V *under* n	1064	10881	9.78	COME	0.768	0.724	0.745
V *with* n	3087	105496	2.93	DEAL	0.704	0.509	0.591
					0.777	**0.531**	**0.612**

3.1 A frequency-ranked type-token VAC profile

The sentences extracted using the procedure described above produced verb type distributions like the following one for the **V *across* n** VAC:

COME	628				
WALK	243				
RUN	202	…			
CUT	198	VEER	3		
…		RIP	3	…	
		…		BUMP	1
				HARE	1
				WHIR	1

These distributions appear to be Zipfian, exhibiting the characteristic long-tail in a plot of rank against frequency. Zipf's law, like other power-law distributions, is most easily observed when plotted on doubly logarithmic axes, where the relationship between log (rank order) and log (frequency) is linear. The advised method to do this is via the (complementary) cumulative distribution (Adamic & Huberman 2002). We generated logarithmic plots and linear regressions to examine the extent of this trend using logarithmic binning of frequency against log cumulative frequency. The binning allows us to select and illustrate an example verb type from each frequency band. Figure 10a shows such a plot for verb type frequency of the **V *across* n** construction; Figure 10b shows the same type of plot

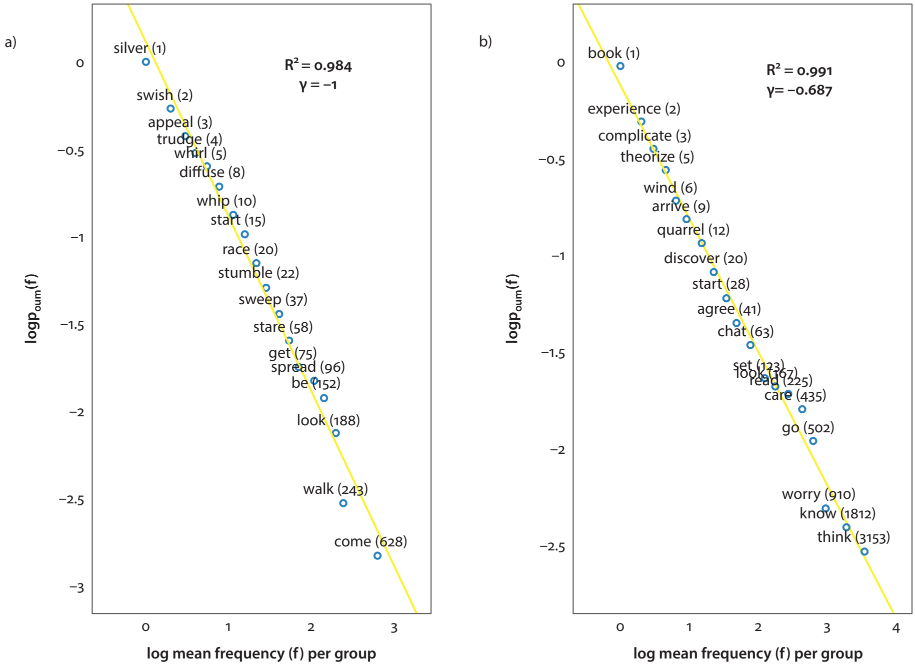


Figure 10. Verb type distributions for (a) **V *across* n** and (b) **V *about* n**

for verb type frequency of the **V *about* n** construction. Both distributions produce a good fit of Zipfian type-token frequency with $R^2 > 0.98$ and slope (γ) around 1. Inspection of the construction verb types, from most frequent down, also demonstrates that the lead member is prototypical of the construction and generic in its action semantics (COME for **V *across* n**; TALK for **V *about* n**).

Since Zipf's law applies across language phenomena, the Zipfian nature of these distributions is potentially trivial. But they are more interesting if the company of verb forms occupying a construction is selective, i.e. if the frequencies of the particular VAC verb members cannot be predicted from their frequencies in language as a whole. We measure the degree to which VACs are selective like this using a chi-square goodness-of-fit test and the statistic '1-tau' where Kendall's tau measures the correlation between the rank verb frequencies in the construction and in language as a whole. Higher scores on both of these metrics indicate greater VAC selectivity. Another useful measure is Shannon entropy for the distribution. The lower the entropy, the more coherent the VAC verb family.

3.2 Determining the contingency between verbs and VACs

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 should be more readily acquired. The measures of contingency adopted here are (1) 'faithfulness' – the proportion of tokens of total verb usage that appear in this particular construction (e.g. the faithfulness of GIVE to the ditransitive is approximately 0.40; that of LEAVE is 0.01) and (2) 'directional mutual information' (MI word → construction: GIVE 16.26, LEAVE 11.73 and MI construction → word: GIVE 12.61 LEAVE 9.11), an information science statistic that has been shown to predict language processing fluency (e.g. Ellis et al. 2008; Jurafsky 2003).

Table 2 lists these contingency measures for the 20 most frequent verbs occupying the **V *across* n** VAC. It also shows the top 20 verbs ordered by the contingency measure MI_{wc} (mutual information in the direction word to construction) and the top 20 ordered according to total corpus frequency (i.e. not just within the VAC). So the top 5 most frequent verbs in **V *across* n** are COME, WALK, RUN, CUT and LOOK. But when the strength of association between verb and construction is considered, the top 5 are SCUD, FLIT, SLANT, SCUTTLE and SKID

Table 2. Top 20 verbs found in the **V *across* n** construction in the BNC

Verb	Constr. freq.	Corpus freq.	Faith.	MI word→ constr	MI constr→ word	Top 20 by MIwc	Top 20 by corpus freq
COME	628	143580	0.004	15.607	10.837	SCUD	BE
WALK	243	19994	0.012	17.081	15.155	FLIT	GO
RUN	202	38688	0.005	15.862	12.984	SLANT	GET
CUT	198	17759	0.011	16.957	15.202	SCUTTLE	SEE
LOOK	188	108373	0.002	14.273	9.908	SKID	COME
BE	152	4090106	0.000	8.728	–0.875	SPRAWL	LOOK
GO	139	224168	0.001	12.788	7.375	TRAMP	PUT
MOVE	136	37573	0.004	15.334	12.498	SCURRY	WORK
LEAN	120	4464	0.027	18.227	18.464	FLICKER	CALL
SPREAD	96	5714	0.017	17.548	17.429	STRIDE	START
GET	75	211788	0.000	11.980	6.649	SPRINT	RUN
FALL	66	26023	0.003	14.821	12.514	SKIM	SET
STARE	58	7573	0.008	16.415	15.890	STUMBLE	MOVE
LAY	55	15799	0.003	15.278	13.691	DIFFUSE	PLAY
STRETCH	55	4446	0.012	17.107	17.350	LEAN	LIVE
TRAVEL	51	8290	0.006	16.099	15.443	FLASH	MEET
REACH	50	22300	0.002	14.643	12.559	SPLASH	CARRY
SET	45	38630	0.001	13.698	10.822	HOP	SIT
STRIDE	44	1049	0.042	18.868	21.195	CRAWL	FALL
LIE	44	13190	0.003	15.216	13.890	SPREAD	REACH

(each of which are of relatively low frequency, both in the VAC and the BNC as a whole).

Table 3 shows these same data for **V *about* n**. TALK, THINK, BE, KNOW and WORRY are the most frequent verbs, while REMINISCE, WORRY, TALK, RAVE and ENTHUSE are most strongly associated (in Mutual Information terms) with the VAC. The intersection of these two orderings (VAC verb frequency and VAC verb contingency) is the overall frequency of a verb in the corpus as a whole, shown in the final columns of Tables 2 and 3. In Section 4 we examine speaker knowledge of verbs in constructions and will consider the effect of VAC verb frequency, VAC verb contingency and verb corpus frequency upon usage.

Table 3. Top 20 verbs found in the **V *about* n** construction in the BNC

Verb	Constr. freq.	Corpus freq.	Faith.	MI word→ constr	MI constr→ word	Top 20 by MI_{wc}	Top 20 by corpus freq
TALK	3832	28867	0.133	16.122	15.870	REMINISCE	BE
THINK	3153	142884	0.022	13.533	10.974	WORRY	SAY
BE	2827	4090106	0.001	8.537	1.138	TALK	GO
KNOW	1812	177192	0.010	12.424	9.554	RAVE	GET
WORRY	910	5822	0.156	16.358	18.416	ENTHUSE	MAKE
SAY	721	314459	0.002	10.267	6.570	GENERALISE	SEE
BRING	712	42271	0.017	13.144	12.342	GENERALIZE	KNOW
HEAR	604	34142	0.018	13.214	12.721	COMPLAIN	TAKE
FORGET	556	11774	0.047	14.631	15.673	FRET	COME
WRITE	517	38144	0.014	12.830	12.176	FUSS	THINK
GO	502	224168	0.002	10.233	7.024	SPECULATE	GIVE
FEEL	482	57807	0.008	12.129	10.876	GOSSIP	LOOK
CARE	435	7607	0.057	14.907	16.579	CARE	FIND
ASK	429	57431	0.007	11.971	10.726	GRUMBLE	TELL
COMPLAIN	427	4206	0.102	15.735	18.262	ENQUIRE	PUT
LEARN	347	18701	0.019	13.283	13.658	INQUIRE	MEAN
SPEAK	320	23910	0.013	12.812	12.832	CHAT	FEEL
FIND	247	95330	0.003	10.443	8.468	FORGET	ASK
READ	225	21154	0.011	12.480	12.677	MOAN	HOLD
WONDER	170	11457	0.015	12.961	14.042	JOKE	BRING

3.3 Identifying the meaning of verb types occupying the constructions and constructing a semantic graph/network

Our semantic analyses use WordNet, a distribution-free semantic database based upon psycholinguistic theory which has been in development since 1985 (Miller 2009). WordNet places words into a hierarchical network. At the top level, the hierarchy of verbs is organised into 559 distinct root synonym sets ('synsets', such as 'move1' expressing translational movement, 'move2' movement without displacement, etc.) which are then split into over 13,700 verb synsets. Verbs are linked in the hierarchy according to relations such as hypernym [verb Y is a hypernym of the verb X if the activity X is a (kind of) Y (*to perceive* is an hypernym of *to listen*], and hyponym [verb Y is a hyponym of the verb X if the activity Y is doing X in some manner (*to lisp* is a hyponym of *to talk*)]. Various algorithms to determine the semantic similarity between WordNet synsets have been developed which consider the distance between the conceptual categories of words, as well as considering the hierarchical structure of the WordNet (Pedersen et al. 2004). Polysemy is a significant issue when working with lexical resources such as WordNet, particularly when analysing verb semantics. For example, in WordNet the lemmas MOVE, RUN and GIVE used as verbs are found in 16, 41 and 44 different synsets, respectively.

In order to analyse VAC meaning patterns, we build semantic networks where the nodes/vertices are the more frequent verb types extracted from the VAC distribution and the edges/links between these nodes indicate some kind of semantic relation. First we construct a similarity matrix consisting of the WordNet Path Similarity scores for each of the pairs of verbs in the matrix. This ranges from 0 (no similarity) to 1 (items in the same synset). There is an extra step to arrive at a similarity score for two verb lemmas, e.g. THINK and KNOW, because WordNet similarity measures work on senses (synsets) and not lemmas. The lemma THINK occurs in 13 different synsets and KNOW in 11. Without carrying out word sense disambiguation to determine which sense of THINK to compare with which sense of KNOW, we calculate scores for each of the 143 possible synset pairs and use the maximum value. For THINK and KNOW this value of path similarity is 0.5 (using the path similarity measure) and results from the synset pair 'remember#v#1' and 'know#v#11' (i.e. the distance in WORD from the first synset for the verb form of REMEMBER and the 11th verb synset for the verb form of KNOW). Next we select a threshold value for the inclusion of an edge between nodes.

An example network for **V *about* n** is shown in Figure 11. The width of the edge (its weight) represents the similarity score. The graph is undirected because the similarity scores are symmetrical. Inspection reveals two major groupings: 1. TALK, WRITE, SPEAK, ASK (with SAY loosely attached) and 2. FIND, LEARN, HEAR, THINK, KNOW. The *COBUILD Grammar Patterns* entry for **V *about* n** (Francis et al. 1996: 146–147) identifies a 'TALK' group, a 'THINK' group and a 'LEARN' group, but this categorisation came from qualitative concordance analysis and introspection. The advantage of our automated methods is that they allow more objective quantitative measurement of the semantic cohesion of the meaning space of VACs using network measures such as network density, average clustering, and degree centrality. We are also implementing and evaluating techniques to identify communities of meaning within the large networks (Ellis et al. 2013; O'Donnell et al. 2012).

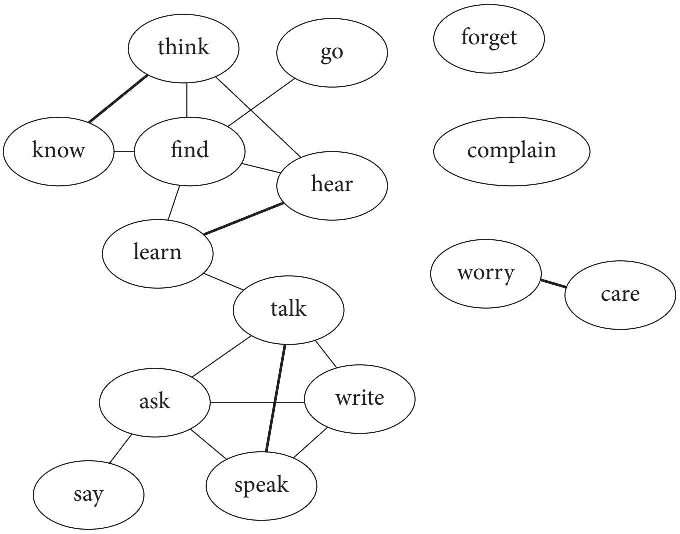


Figure 11. Semantic similarity network for the top 15 verbs in **V *about* n** using the WordNet Path Similarity measure

4. VACs in the mind: Native speaker and learner evidence

The corpus analyses above have highlighted central structural and distributional properties of a range of VACs in language use. Our next goal was to test whether these properties are the same for VAC representations in the minds of language users. This involved addressing the following questions: do the corpus findings mirror what speakers know about verbs in constructions or are speaker mental representations different from usage? Does the verb/construction knowledge of native speakers differ from that of non-native speakers?

In order to address these questions, we had English native speakers and advanced English language learners (L1 German) complete the same type of generative free association task. We designed a series of experiments involving a web-based survey in which we presented participants with VAC frames such as *she* ___ *about the …* or *it* ___ *across the…* and asked them to type the first word that came to mind to fill the slot. For each VAC, we recorded the range of verbs that subjects generated as well as their speed of access. In the first two experiments we collected responses on 20 VACs from 276 native speakers (Experiment 1) and 276 German learners (Experiment 2). We lemmatised the lists of responses by verb type and ordered them by token frequencies. We then compared the results of the experiments with the results from the previous BNC analyses (see Section 3). We also compared Experiment 1 and Experiment 2 against each other to determine how closely native-speaker and learner knowledge of the selected VACs are related and in what ways the two groups differ in terms of verb selection preferences. The following sections discuss our findings for two verb-argument constructions: **V *across* n** and **V *about* n**.

4.1 Native speaker and learner verb preferences vs. corpus findings

For all 20 VACs, including **V *across* n** and **V *about* n**, we observed strong correspondences between the frequencies of verb exemplars in natural usage (BNC data) and native speakers' free associations to particular frames (see Ellis et al. 2014). Figures 12 and 13 illustrate the correlations between the responses from Experiments 1 and 2 and the corpus data for **V *across* n** and **V *about* n** in scatter plots.

The y-axis shows the logarithmic frequency of the verb type in the speakers' responses in the two experiments. The x-axis shows the logarithmic frequency of the verb type in the VAC from the search of the BNC. For example, in Figure 12, the native speaker responses for **V *across* n** show RUN, COME and SWIM as the most frequent responses, with frequencies of 88, 56 and 43 respectively. In the BNC search for the same VAC, these three verbs have ranks and frequencies as follows: RUN (3 – 202), COME (1 – 628) and SWIM (46 – 19). A perfect correspondence between frequency in speaker responses and VAC verb frequency would place all verbs on the diagonal. Items in the right (lower) corner of the plot are markedly less frequent in the speaker responses than their relative frequency in the corpus distribution, e.g. CUT *across* (corpus – rank: 4, frequency: 198; L1 English – rank: 98, freq: 1). Conversely, verbs in the (top) left region of the plot are relatively much more frequent in the speaker responses than in the corpus VAC distribution, e.g. SKATE *across* (corpus – rank: 362, freq: 1; L1 English – rank: 12, freq: 7). The size of the word in the scatter plots indicates its overall frequency in the BNC.

Each plot also shows the correlation coefficient *r* for the comparison.[3] It is interesting to note that for **V *across* n** the German L1 speakers' responses match the corpus distribution more closely (*r*=0.884) than do those of the native English speakers (*r*=0.633). Regional variation in English could be at least a partial possible explanation for this finding. While our native speaker participants are almost exclusively speakers of American English, British English is the norm that the majority of our advanced German learners wish to approximate to. This may be why their responses are more in line with VAC distributions in the British National Corpus. In order to address this issue and confirm this hypothesis, we have started to collect survey data from native speakers of British English.

[3] Pearson's correlation coefficient: r=1 perfect positive correlation; r=0 no correlation; r=−1 perfect negative correlation.

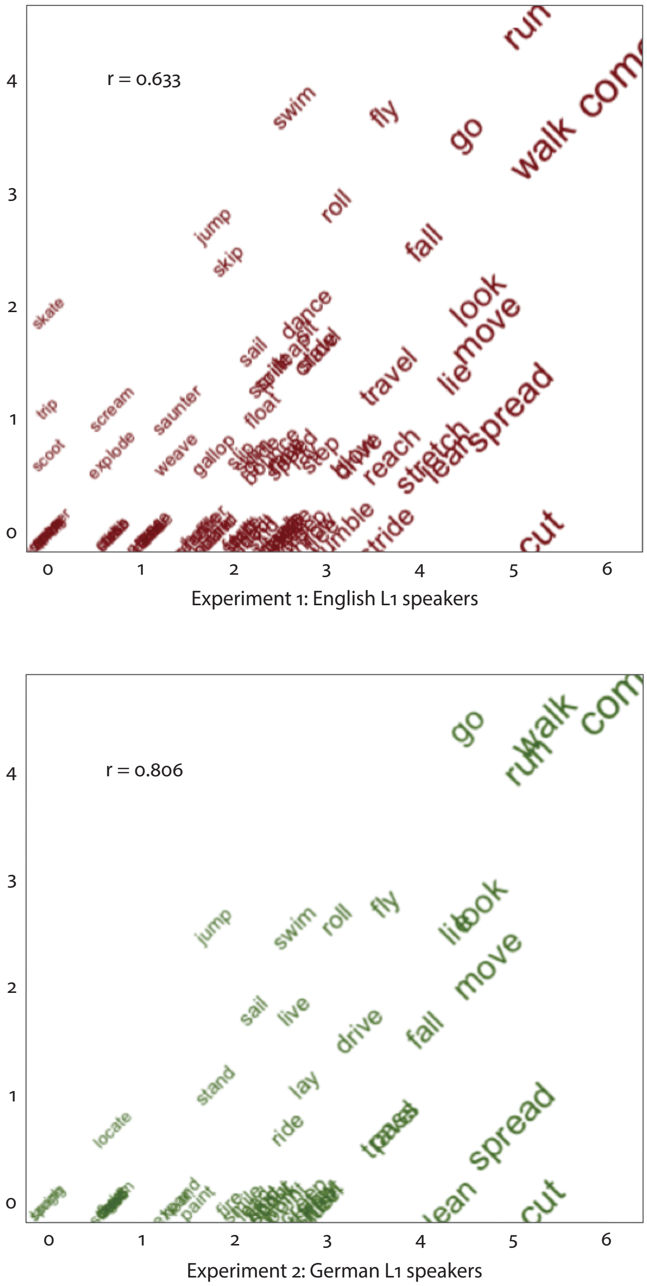


Figure 12. V *across* n, comparison of native speaker (Experiment 1) and L1 German speaker (Experiment 2) responses against corpus data (VAC frequencies). x-axis = log verb in VAC frequency in BNC; y-axis = log verb in speaker responses

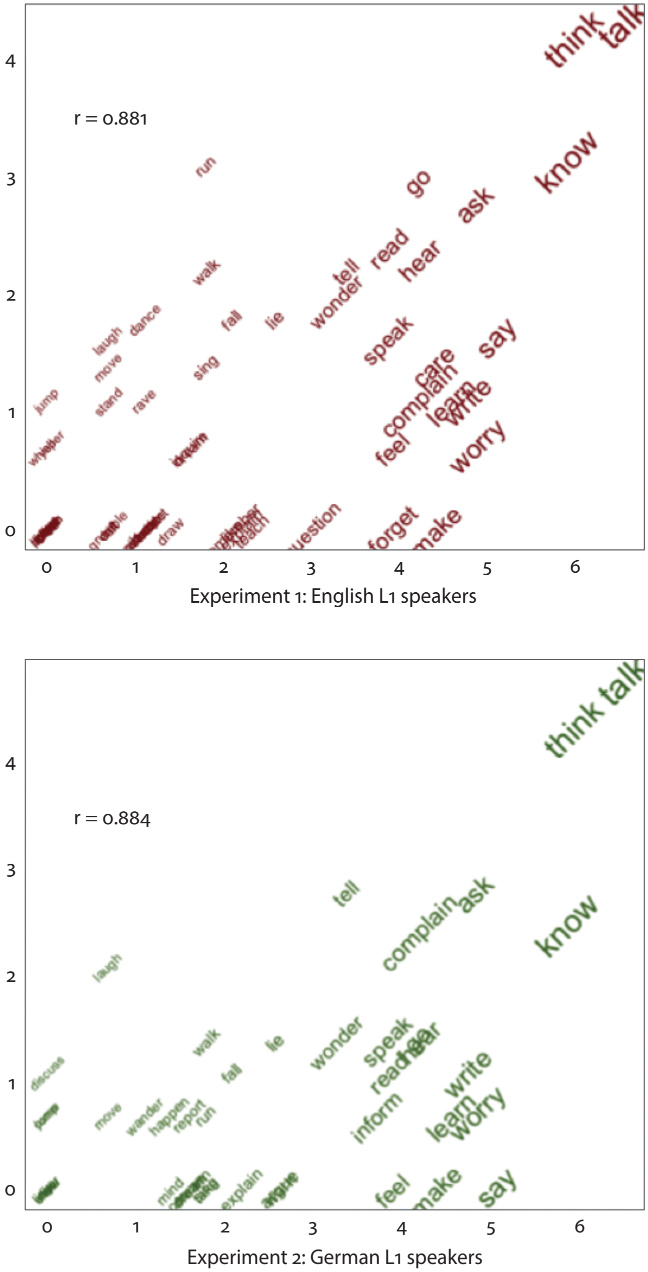


Figure 13. V *about* n: comparison of native speaker (Experiment 1) and L1 German speaker (Experiment 2) responses against corpus data (VAC frequencies). x-axis = log verb in VAC frequency in BNC; y-axis = log verb in speaker responses

For the **V *about* n** VAC we did not observe a similar effect. The correlation coefficients for the speaker/BNC comparisons are almost identical for the learner (*r*=0.884) and native speaker respondents (*r*=0.881) (see Figure 13). Despite the very abstract prompts they were given in the form of bare frames, native speaker and non-native speaker subjects generate verb-preposition clusters that are similar to those found in the BNC analyses. This implies that both groups of respondents have intuitions about verbs in this construction that are very much in line with the frequency distributions found in actual language usage. There are only a few verbs which are comparatively more frequent in the BNC VAC data than in the English and German speaker data, including SAY, WORRY, LEARN, and FEEL. On the other hand, a number of low-frequency verbs (small font type in the scatter plots in Figure 13) are relatively more common in the speaker responses than in the BNC VAC frequency list. The verb LAUGH belongs to this group, as do the motion verbs JUMP, MOVE, WALK, RUN, and DANCE.

4.2 Learner vs. native speaker verb preferences

In addition to the comparisons of BNC findings with native speaker survey data (Experiment 1) and BNC findings with German learner survey data (Experiment 2), we also analysed for the selected VACs how similar or different the native speaker and learner responses were. Figures 14 and 15 summarise the results of these comparisons for the **V *across* n** and **V *about* n** constructions. Lists of the 20 most frequent verbs in Experiments 1 and 2 are given in Table 4 (**V *across* n**) and Table 5 (**V *about* n**). Verbs that appear in both lists (native speaker and German learner) are displayed in italics.

The first thing we observe for the **V *across* n** construction is a strong correspondence between native speakers' and learners' free associations to the VAC frame. The *r*-value for this VAC is 0.789, and 11 out of the top-20 verbs are shared across lists. Both experiment groups most frequently respond to the frame with verbs of physical motion such as RUN, COME, WALK, GO, FLY and SWIM. The verbs ROLL and FALL, which also occur in both lists in Table 4, may have been picked up from the survey instructions (the examples *it rolls down the…* and *it fell down the…* were given at the beginning of the online questionnaire to illustrate the procedure). As their font size in Figure 14 indicates, RUN, COME, WALK and GO also occur very frequently in the BNC. FLY and SWIM are not quite as frequent in the BNC data (smaller font size) and occur comparatively more often in the native speaker than in the learner responses (see their position in the plot in Figure 14 and the response counts in Table 4). Other verbs produced considerably more often by native speaker informants than German learners include SKIP (rank 10 in native speaker list), SKATE (rank 12), SIT (rank 14) and SLIDE (rank 19). None of these verbs appears in the top-20 verb list based on learner responses. For SKIP, SKATE, SIT and SLIDE we observed low BNC token frequencies but high faithfulness scores (see values in Table 2).

Table 4 also lists a few verbs that are produced repeatedly by the L1 German participants but not at all (or very infrequently) by the 276 native speakers. Among this group of verbs are LIE (rank 10 in the learner responses), DRIVE (rank 15), TRAVEL (rank 18) and SPREAD (rank 19). These verbs show high BNC frequencies, as illustrated by their font sizes in Figure 14 (see especially SPREAD and LIE), but lower contingency scores for this particular VAC. These findings suggest that the native speakers who participated in the survey produced verbs which are more strongly associated with the **V *across* n** frame in natural production data while the German learners essentially picked up on high frequency verbs.

In the data we collected for the **V *about* n** construction, the correlations between native speaker and German learner responses are even more pronounced than for **V *across* n**. Figure 15 and Table 5 summarise the results on verb selections by both groups of respondents. We found a very high *r*-value of 0.937 for this comparison, and 17 of the 20 most frequently produced verbs for this VAC are shared among native speaker and learner lists. The participants in both groups most commonly responded to this VAC frame with verbs of communication and cognition, including TALK, ASK, TELL, SPEAK, HEAR, THINK, KNOW and WONDER. The number one verb in both lists is TALK. Of the 276 German learners, 117 typed in a form of this verb, either as a response to the *it __ about the…* frame or to *she __ about the….* Apart from the verbs that were also produced by the native speaker informants, a number of our German learners responded with the verb COMPLAIN to this VAC frame. COMPLAIN is not used by the native speakers in the survey but it has a fairly high frequency in the BNC. The same applies to WRITE, which is also covered in the German learner top-20 but not in the native speaker list (see Table 5).

Finally, the verb DISCUSS presents a very interesting case. Three of our German respondents completed the *she/he __ about the…* frame with the form *discuss* or *discussed.* A prescriptive grammar of English would consider this type of use ungrammatical. It is, however, debatable whether we are here dealing with a learner error or a creative new usage of language that has also been observed in spoken English as a Lingua Franca (Mauranen 2012) and in postcolonial or 'outer circle' varieties of English, such as Indian English (Mukherjee 2009: 123). Of course, in the case of German learners, *discuss about* could be interpreted as a transfer phenomenon from the first language (German 'diskutieren über') but given the evidence from other sources, we suspect that this is a wider-ranging phenomenon.

We may be observing the development of a new prepositional verb, formed in analogy with *speak about* or *talk about*.[4] In fact, there are more than 20 instances of *discussed about* and a dozen instances of *discuss about* in the current (20 Nov 2011) version of the Corpus of Contemporary American English (COCA), including one from a 2009 ABC Primetime interview with Barack Obama in which the US president says: "It doesn't come from the evidence-based care and changes in reimbursement that I've already discussed about, […]".

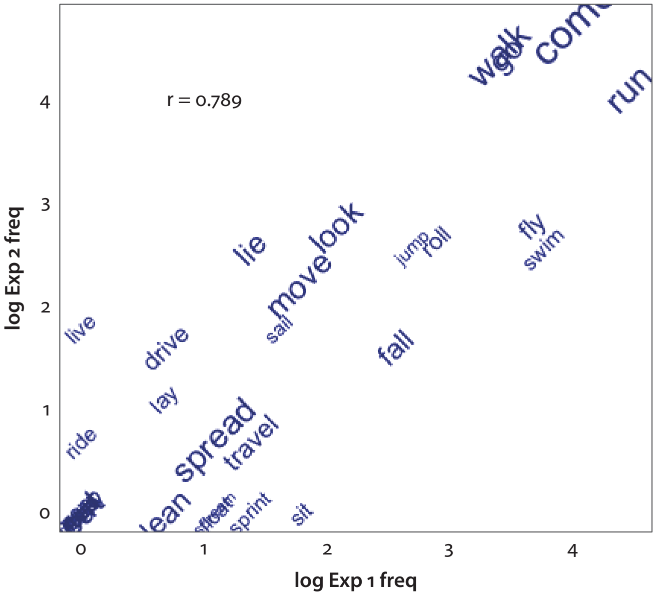


Figure 14. V *across* n: comparison of native speaker (Experiment 1) and German learner responses (Experiment 2)

Table 4. V *across* n, top-20 verbs in native speaker and German learner responses

Rank	Native speakers		German learners	
1	*RUN*	88	*COME*	113
2	*COME*	56	*WALK*	85
3	*SWIM*	43	*GO*	82
4	*FLY*	40	*RUN*	60
5	*GO*	33	*FLY*	16

(*Continued*)

[4] For further discussion of the role that pattern-based analogies might play in language change see Hunston & Francis (1999: 96–98).

Table 4. (*Continued*)

Rank	Native speakers		German learners	
6	*WALK*	30	*LOOK*	16
7	*ROLL*	18	*ROLL*	14
8	*JUMP*	15	*SWIM*	13
9	*FALL*	13	*JUMP*	13
10	SKIP	11	LIE	13
11	*LOOK*	8	*MOVE*	9
12	SKATE	7	*SAIL*	6
13	DANCE	7	LIVE	6
14	SIT	6	*FALL*	5
15	*MOVE*	6	DRIVE	5
16	CRAWL	5	LAY	3
17	LEAP	5	STAND	3
18	*SAIL*	5	TRAVEL	2
19	SLIDE	5	SPREAD	2
20	SPILL	4	RIDE	2

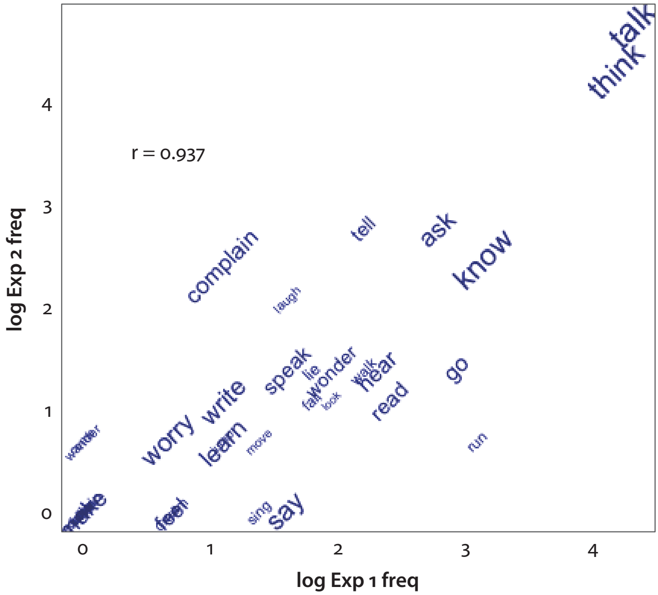


Figure 15. V *about* n: comparison of native speaker (Experiment 1) and German learner responses (Experiment 2)

Table 5. Top-20 **V *about* n** verbs in native speaker and German learner responses

Rank	Native speakers		German learners	
1	*TALK*	74	*TALK*	117
2	*THINK*	65	*THINK*	75
3	*KNOW*	23	*ASK*	16
4	*RUN*	22	*TELL*	16
5	*GO*	19	*KNOW*	12
6	*ASK*	16	COMPLAIN	11
7	*READ*	11	*LAUGH*	8
8	*HEAR*	10	*GO*	4
9	*TELL*	9	*HEAR*	4
10	*WALK*	9	*WALK*	4
11	*LOOK*	7	*WONDER*	4
12	*WONDER*	7	*LIE*	4
13	*FALL*	6	*SPEAK*	4
14	*LIE*	6	*READ*	3
15	DANCE	6	*LOOK*	3
16	SAY	5	*FALL*	3
17	*SPEAK*	5	WRITE	3
18	*LAUGH*	5	DISCUSS	3
19	SWIM	5	*RUN*	2
20	*MOVE*	4	*MOVE*	2

5. Conclusion

This paper has taken a selection of patterns defined in *COBUILD Grammar Patterns 1* (Francis et al. 1996) as a starting point for an in-depth analysis of verb-argument constructions in English. It has discussed some important methodological issues in corpus mining (including questions related to precision/recall) and suggested an innovative approach to making verb construction analyses scalable. It has also addressed questions about the type/token frequencies and semantics of VACs and about native and non-native speakers' knowledge of verbs in constructions. Corpus analyses showed how verbs are distributed across constructions and which preferences constructions have for certain groups of verbs, and vice versa. Major challenges lay in balancing automatic corpus extraction and human intervention, and in defining BNC searches based on COBUILD descriptions that provide accurate and comprehensive results of the VACs in question. Through the development of a cyclical process of corpus searches, precision and recall analyses, and revisions of search strings, we were able to refine our initial searches and increase the amount and accuracy of VAC results retrieved from the corpus.

Through psycholinguistic experiments we gained insights into how entrenched selected VACs are in the native speaker's and in the second language learner's mind. Our findings, based on a set of 20 VACs indicate that native and non-native speakers have a strong constructional knowledge and make selections which, to a large extent, match actual usage patterns. Even rather bare grammatical frames seem to carry fairly specific meanings and trigger semantically related verbs that are also found in the respective VAC in the corpus data. While there are some interesting differences between learners' and native speakers' psychological associations of frames and lexical items, there is a large amount of overlap among the most common responses, at least for the VACs discussed in this paper. The responses of both groups show strong correlations with the data retrieved from the BNC. These findings demonstrate that VACs are psychologically real – not just in the minds of native speakers but also in the minds of advanced second language learners.

One goal of this paper was to make a case for combining insights, tools and techniques from corpus, computational and psycholinguistics. We believe that this kind of interdisciplinary work can lead to important findings that are of relevance to linguistic description, second language acquisition theory and pedagogical practice (cf. McEnery & Hardie 2012: 192–223). In the case of the present study, one implication would be that constructions (and phraseology in general) need to be taken more seriously in theory and practice. Our VAC analyses provide additional evidence for the inseparability of lexis and grammar and hence support claims made in Construction Grammar and pattern grammar studies. Our findings on verb-construction associations extend previous studies in terms of coverage, statistical detail and semantic grouping. We are currently working on expanding the analyses to include a larger set of VACs and cover a range of different L1s in the psycholinguistic experiments.

In the conclusion to *Pattern Grammar*, Hunston and Francis say that they "look forward to the development of an automatic pattern identifier" (1999: 272). As far as we know, such a tool is still non-existent, but we think that the suite of tools and series of analytic techniques we have developed in the context of the VAC project brings us a little closer to that goal.

Acknowledgements

The authors would like to thank the following students for their indispensable help with the definition of BNC searches, precision and recall analyses, and distribution of the VAC knowledge surveys: Gin Corden, Katie Erbach, Mary Smith, Danny Tzu-Yu Wu, and Lucy Zhao.

References

Adamic, L.A. & Huberman, B.A. 2002. Zipf's law and the Internet. *Glottometrics* 3: 143–150.

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

Briscoe, E.J., Carroll, J. & Watson, R. 2006. The Second Release of the RASP System. *Proceedings of the COLING/ACL 2006. Interactive Presentation Sessions, Sydney, Australia*, 77–80. DOI: 10.3115/1225403.1225423

Ellis, N.C. & Ferreira-Junior, F. 2009. Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics* 7: 111–139. DOI: 10.1075/arcl.7.05ell

Ellis, N.C., O'Donnell, M.B. & Römer, U. 2013. Usage-based language: Investigating the latent structures that underpin acquisition. *Language Learning* 63(Supp. 1)*:* 25–51. DOI: 10.1111/j.1467-9922.2012.00736.x

Ellis, N.C., O'Donnell, M.B. & Römer, U. 2014. The processing of verb-argument constructions is sensitive to form, function, frequency, contingency, and prototypicality. *Cognitive Linguistics* 25(1): 55–98.

Ellis, N.C., Simpson-Vlach, R. & Maynard, C. 2008. Formulaic language in native and second-language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly* 42(3): 375–396.

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

Goldberg, A.E. 2003. Constructions: A new theoretical approach to language. *Trends in Cognitive Science* 7: 219–224. DOI: 10.1016/S1364-6613(03)00080-9

Goldberg, A.E. 2006. *Constructions at Work. The Nature of Generalization in Language*. Oxford: OUP.

Goldberg, A.E., Casenhiser, D.M. & Sethuraman, N. 2004. Learning argument structure generalizations. *Cognitive Linguistics* 15: 289–316. DOI: 10.1515/cogl.2004.011

Hoffmann, S., Evert, S., Smith, N., Lee, D. & Berglund Prytz, Y. 2008. *Corpus Linguistics with BNCweb – A Practical Guide* [English Corpus Linguistics 6]. Frankfurt: Peter Lang.

Hunston, S. & Francis, G. 1999. *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English* [Studies in Corpus Linguistics 4]. Amsterdam: John Benjamins. DOI: 10.1017/s0022226701001001

Jurafsky, D. 2003. Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In *Probabilistic Linguistics*, R. Bod, J. Hay & S. Jannedy (eds), 39–96. Harvard MA: The MIT Press.

Mason, O. & Hunston, S. 2004. The automatic recognition of verb patterns: A feasibility study. *International Journal of Corpus Linguistics* 9: 253–270. DOI: 10.1075/ijcl.9.2.05mas

Mauranen, A. 2012. *Exploring ELF: Academic English Shaped by Non-native Speakers*. Cambridge: CUP.

McEnery, A. & Hardie, A. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: CUP.

Miller, G.A. 2009. *WordNet – About Us.* Princeton NJ: Princeton University.

Mukherjee, J. 2009. The lexicogrammar of present-day Indian English: Corpus-based perspectives on structural nativisation. In *Exploring the Lexis-Grammar Interface* [Studies in Corpus Linguistics 35], U. Römer & R. Schulze (eds), 117–135. Amsterdam: John Benjamins. DOI: 10.1075/scl.35.9muk

O'Donnell, M.B., Ellis, N.C. & Corden, G. 2012. Exploring semantics in verb argument constructions using community identification algorithms. Paper presented at the Language & Network Symposium: The International Conference on Network Science NETSCI 2012.

Pedersen, T., Patwardhan, S. & Michelizzi, J. 2004. WordNet::Similarity – Measuring the relatedness of concepts. *Proceedings of the Fifth Annual Meeting of the North American Chapter of the Association of Computational Linguistics (NAACL 2004)*, 38–41.

Shanks, D.R. 1995. *The Psychology of Associative Learning*. Cambridge: CUP. DOI: 10.1017/CBO9780511623288

Sinclair, J.M. 2003. *Reading Concordances: An Introduction*. London: Longman.

Sinclair, J.M. 2004. *Trust the Text. Language, Corpus and Discourse*. London: Routledge.

Zipf, G.K. 1935. *The Psycho-biology of Language. An Introduction to Dynamic Philology*. Cambridge MA: The MIT Press.