DE GRUYTER OPEN

ICAME Journal, Volume 38, 2014, DOI: 10.2478/icame-2014-0006

Linking learner corpus and experimental data in studying second language learners' knowledge of verb-argument constructions*

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

Abstract

This paper combines data from learner corpora and psycholinguistic experiments in an attempt to find out what advanced learners of English (first language backgrounds German and Spanish) know about a range of common verb-argument constructions (VACs), such as the 'V about *n' construction (e.g.* she thinks about chocolate a lot*). Learners' dominant verb-VAC associations are examined based on evidence retrieved from the German and Spanish subcomponents of ICLE and LINDSEI and collected in lexical production tasks in which participants complete VAC frames (e.g. 'he ___ about the...') with verbs that may fill the blank (e.g.* talked, thought, wondered*). The paper compares findings from the different data sets and highlights the value of linking corpus and experimental evidence in studying linguistic phenomena.*

1 Introduction: Studying English verb-argument constructions (VACs)

Over the past few years we have seen an increase in studies that highlight the value of combining corpus and experimental evidence in the study of linguistic phenomena (consider, for example, Ellis and Simpson-Vlach 2009; Gilquin and Gries 2009; Wulff 2009; Arppe, Gilquin, Glynn, Hilpert and Zeschel 2010; Römer, O'Donnell and Ellis 2012). Many of these studies utilize corpora of native speaker output (such as the BNC or ICE-GB) to derive frequency data which is then considered in relation to speaker responses or judgments collected in experimental settings. Such studies demonstrate that different types of data

can present converging evidence which helps strengthen research hypotheses. They also show that a combination of data types allows us to ask a wider range of questions and go beyond existing work in corpus linguistics. The present paper discusses the role that second language learner (rather than native speaker) output, as captured in learner corpora, may play in the context of combining different types of empirical evidence in linguistic analysis.

We investigate how learner corpora and experimental data complement each other in providing insights into second language learner knowledge of 19 different verb-argument constructions (VACs), including the 'V *against* n' and the 'V *like* n' constructions (as exemplified by *she leaned against the wall* or *he ran like the wind*). We are interested in finding out which verbs learners of two different first languages (L1 German and L1 Spanish) most commonly associate with a particular VAC and whether and how their verb-construction associations are different from those of native speakers of English. Corpus sources are the German and Spanish subsets of the International Corpus of Learner English (ICLE) and of the Louvain International Database of Spoken English Interlanguage (LINDSEI). In addition to the corpus data, we draw on data collected in lexical production tasks which ask participants to complete bare VAC frames (e.g. 'She ____ against the...') with verbs that may fill the blank (e.g. *pushed*, *leaned*, *ran*). A comparison of the results from the corpus and survey data analysis helps us assess the contributions of these different types of data to a better understanding of learner VAC knowledge.

The context of this study is a collaborative project which aims to build an inventory of a large number of VACs in English language use. The project takes constructions identified and discussed in the *COBUILD Grammar Patterns: Verbs* volume (Francis, Hunston and Manning 1996) as a starting point for a systematic analysis of VACs in the British National Corpus (BNC; see Römer, O'Donnell and Ellis forthcoming for a description of the BNC VAC extraction). With the help of psycholinguistic experiments, the project also studies native-speaker and learner associations of verbs and the selected VACs. Comparisons of the results from the BNC analyses and the knowledge experiments allow us to determine in what ways and to what extent speakers are affected by verb distributions in the input (see Ellis, O'Donnell and Römer 2013, forthcoming). Central findings of the study include: verb-argument constructions are (1) Zipfian in their type-token distribution with one verb type accounting for the lion's share of all VAC tokens, (2) selective in their verb form occupancy, and (3) coherent in their semantics (Ellis, O'Donnell and Römer 2013). The psycholinguistic experiments and comparisons of BNC and survey data indicated that both L1 and L2 English speakers have construction knowledge and that speaker verb

production in VACs is influenced (1) by the token frequency of verbs in VACs in general language usage, (2) by the faithfulness of verbs to particular VACs in usage, and (3) by the centrality of the verb meaning in the VAC's semantic network in usage (Ellis, O'Donnell and Römer forthcoming).

So far, the only type of evidence that the project has collected and evaluated to investigate what second language learners know about verbs in VACs is data from free association tasks in which learners of English generate the first word that came to mind to fill the verb slot in 20 sparse VAC frames such as 'She ____ against the...' (see Römer, O'Donnell and Ellis submitted). The present study expands on this work by including (1) a richer type of experimental data and (2) additional VAC evidence retrieved from spoken and written learner corpora. One aim of our study is to determine which verbs L1 German and L1 Spanish learners of English most commonly associate with a particular VAC and whether and how their verb-VAC associations differ from those of L1 English speakers. Another aim is to address the following methodological questions: How do learner corpus and experimental data complement each other in providing insights into L2 learners' knowledge of frequent VACs? Are we dealing with converging or diverging evidence? What are the potential benefits of linking different types of data in the study of a second language acquisition phenomenon? We believe that finding answers to these questions will be important as we work towards more mixed-methods approaches in corpus linguistics which combine various types of empirical evidence.

After a description of the data types and methods of data collection (Section 2), the paper will present selected results from the analyses of the corpus and experimental evidence on L2 learner VAC knowledge (Sections 3 and 4). In our comparison of corpus and survey results (Section 5), we will consider the usefulness of the two selected types of data in the given context and address the question of whether they present us with converging or diverging evidence on learner knowledge of VACs. The paper closes with a discussion of how a combination of learner corpus and experimental data may lead us to a more comprehensive understanding of the extent to which VACs are entrenched in the second language learner's mind and lists tasks for future work in this area.

2 Data and methods

Two types of data were collected and analyzed to help us better understand what advanced English language learners know about a selection of common VACs and which verbs they associate most strongly with those constructions: (1) data from spoken and written learner corpora, and (2) data produced in psycholin-

guistic experiments. These two data types differ in terms of naturalness, or in the extent to which they mirror actual communicative practices. According to Gilquin and Gries (2009), prototypical corpora such as the BNC may be considered the most natural type of linguistic data, followed by less prototypical corpora consisting of authentic texts that have, however, not been produced in natural communicative settings, which would include most learner corpora. The authors rank psycholinguistic experiments "requiring subjects to do something with language they usually do not do" (2009: 5) among the least natural types of linguistic data. We will refer back to this issue in our discussion of the usefulness of the selected data sources in Section 5.

From the sources further described below, data was collected for the following verb-argument constructions, all selected from Chapter 2 of the *COBUILD Grammar Patterns: Verbs* volume (Francis, Hunston and Manning 1996): 'V *about* n', 'V *across* n', 'V *after* n', 'V *against* n', 'V *among* n', 'V *around* n', 'V *as* n', 'V *between* n', 'V *for* n', 'V *in* n', 'V *into* n', 'V *like* n', 'V *of* n', 'V *off* n', 'V *over* n', 'V *through* n', 'V *towards* n', 'V *under* n', and 'V *with* n'.

2.1 Learner corpus analysis

For the first part of our VAC knowledge analysis, we collected data from subsections of the International Corpus of Learner English (ICLE) and the Louvain International Database of Spoken English Interlanguage (LINDSEI). The texts in ICLE are argumentative essay responses to supplied topics (e.g. the effect of technology on imagination, the value of university degrees) written by advanced undergraduate EFL learners from sixteen different first language backgrounds (Granger, Dagneaux, Meunier and Paquot 2009). For the current analysis, we selected the Spanish native speaker (198,109 words) and German native speaker sections (236,095 words) of the corpus. LINDSEI contains informal interviews between interviewers and EFL undergraduate university students from eleven first language backgrounds; the majority of students' language proficiency was rated as high intermediate (Gilquin, De Cock and Granger 2010). For this study, we selected the Spanish native speaker (63,889 words) and German native speaker (86,072 words) sub-sections of the corpus. For both of these, we excluded interviewer speech from the word count and analysis. Corpus collection tasks for LINDSEI include a warm-up where learners choose from among three topics to discuss freely (a lesson they have learned, a country they have visited, a film they have seen), followed by an informal discussion about the chosen topic that constitutes the main part of the spoken interaction. Finally, learners look at four pictures and make up a story to describe what they see.

To extract the desired VAC data from ICLE, we used the interface that accompanies the corpus to select essays written by Spanish and German examinees. Since ICLE is part-of-speech tagged, we were able to use verb tags to extract instances of verbs directly followed by one of the nineteen prepositions listed above. We used a search string that included each of the five verb tags used in ICLE (Vbe, Vdo, Vhave, Vlex, Vmod), plus a preposition (e.g., *about, across*), such that a search string may read '<Vbe> *across*'. The resulting concordance lines of each search were exported to Excel and manually filtered for true instances of the VAC. Raw hits were filtered to ensure that the second element of the search string (*about*, *across*, etc.) was used as a preposition, and not, for example as an adverb, as in *the stepmother of Theresa was about only twenty-five* (ICLE German), and to ensure that the preposition was followed by a noun or noun group (hits like *it's rarely talked about.* (ICLE German) were eliminated).

For LINDSEI, we again used the CD-ROM search interface to identify German and Spanish learner interviews, limit the text selection to learner speech only, and save the output as text files. Since LINDSEI is not tagged for parts of speech, we identified relevant instances of the selected VACs by carrying out preposition searches (*about, across*, etc.) in WordSmith Tools (6.0), sorting the context to the left of the search term, and manually filtering the resulting concordances for true hits of the VACs.

This data extraction process yielded raw and filtered datasets of largely varying sizes. Raw hits for ICLE subsets ranged from three to 1,123, and filtered from one to 647. Hits for LINDSEI subsets ranged from one to 1,399 (raw) and from zero to 344 (filtered). We will provide an overview of the resulting instances per VAC and sub-corpus in Section 3 below.

2.2 Psycholinguistic experiments

The second type of data used in our study comes from lexical production tasks in which participants complete VAC frames (e.g. 'she ___ against the...') with verb forms that may fill the blank (e.g. *pushed*, *leaned*, *ran*). Previous research asked respondents to generate the first verb that came to mind in a particular VAC frame (Römer, O'Donnell and Ellis forthcoming, submitted; Ellis, O'Donnell and Römer forthcoming). This type of free association task is useful for generalizing across language users, but it does not tap the depth or bounds of VAC knowledge in particular users. The current study expands on previous experiments by using a verbal fluency test that asks native speaker, L1 German, and L1 Spanish participants to generate as many verbs as possible for a given VAC frame over a span of 60 seconds. The results of this type of fluency test

provide evidence about the typicality of verbs in a VAC, as more typical verbs are likely to be produced earlier, and by more participants, than less prototypical ones (Gruenewald and Lockhead 1980).

By means of a web-based Qualtrics survey (www.qualtrics.com), participants were presented with 20 bare VAC frames of the type 'he/she/it ___ PREP the...', for example 'she ___ against the…' and 'it ___ towards the...' (see Figure 1 for an example of an included survey prompt). We recruited students from universities in the US, Germany and Spain to participate in the survey and offered them an Amazon gift card (worth five USD or five EUR, respectively) to compensate them for their time. 99 American English native speakers, 94 advanced L1 German and 96 L1 Spanish learners of English completed the survey. Participants were instructed to type in the first verbs they could think of that could fill the gap and to press the enter key after each verb. They were informed that the one minute countdown begins when they start typing the first verb and that they can take breaks between questions (when a new VAC frame appears, before they enter any words). They then saw 20 sentence frames, each one for 60 seconds. We recorded participant responses and the time they took between responses. Responses from each of the three participant groups were lemmatized using the Natural Language Toolkit (NLTK, Bird et al. 2009). Frequency-sorted versions of the lemmatized verb lists were compared against each other and against the verb lists that resulted from the ICLE and LINDSEI analyses of the same VACs.

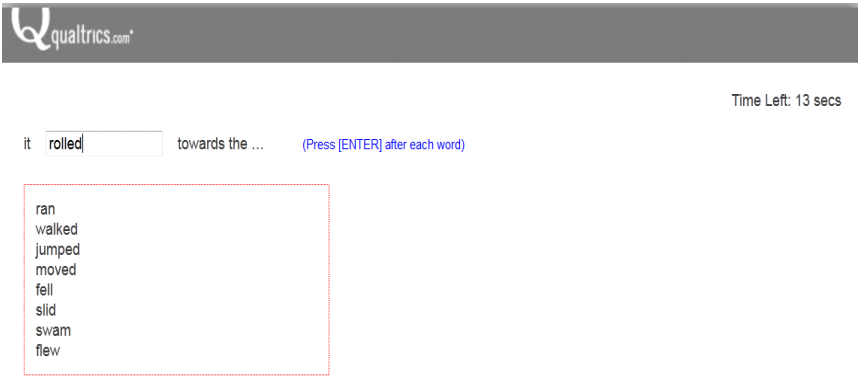

Figure 1: Example survey prompt, designed to trigger verb responses for the 'V towards *n' VAC*

3 Results (I): Learner corpus evidence on second language learner VAC knowledge

This section presents results from our ICLE and LINDSEI analyses, focusing on L1 Spanish and L1 German learner knowledge of VACs. We will first present an overview of learner corpus results for all selected VACs, including verb type and token information, and listing the most common verb for each construction. We will then focus on two VACs ('V *about* n' and 'V *with* n') for a more in-depth discussion of learners' verb selection preferences, highlighting differences across corpora and L1 learner groups.

Table 1 presents an overview of type and token numbers for the nineteen selected VACs across corpora and first language groups. This overview of corpus data highlights interesting patterns in frequency of VACs. Overall, inconsistency in token numbers (ranging from zero to 647) underscores the need for examining this phenomenon using multiple data types. Some constructions are very infrequent across both language backgrounds and corpora. An example of this is 'V *among* n', where the token counts range from zero to five for all four data sets. For other VACs, such as 'V *like* n' and 'V *of* n', there appears to be a register effect. The former is much more common in the spoken corpus subsets, while the latter appears more frequently in the written ones. There is only a handful of VACs that display robust token counts across corpora and language backgrounds. 'V *about* n', 'V *for* n', 'V *in* n', and 'V *with* n' fall in this group. Two of these, 'V *about* n' and 'V *with* n', will be discussed in more detail below.

Table 2 provides the top most frequent verb in each VAC across the four learner corpus data sets. For many VACs, it would be misleading to talk about a 'lead verb' because of low overall token numbers and hence low token numbers of the most frequent verbs. Number one verbs that have token frequencies of less than five in a data set hence appear in parentheses in Table 2. The label 'n/a' (not applicable) was used when there were no hits for a VAC or only single occurrences of individual verb types.

For the majority of VACs, low token frequencies in ICLE and LINDSEI, especially in the Spanish subsets, make it impossible to identify lead verbs. This applies to 'V *across* n', 'V *after* n', 'V *against* n', 'V *among* n', 'V *around* n', 'V *between* n', 'V *like* n', 'V *off* n', 'V *over* n', 'V *through* n', 'V *towards* n', and 'V *under* n'. This scarcity of hits calls for a consultation of additional data sources that may help us obtain a more complete picture of verb-VAC associations. For the remaining VACs with larger numbers of instances, we observe that number one verbs are shared across most data sets. German and Spanish learners most often use forms of the verb TO BE in the 'V *in* n' VAC and forms of the

verb TO THINK in 'V *of* n', independent of whether they are producing written or spoken text. TALK and WORK are the dominant verbs in the 'V *about* n' and 'V *as* n' constructions, respectively.

For two other VACs, the type of corpus and the kind of text included in it appear to determine the most common verb. For 'V *with* n', DEAL is the lead verb in both ICLE subsets (often used in descriptions of literary works, e.g. *the play deals with a family*, ICLE Spanish), but not in LINDSEI where WORK and BE are most frequently used in this VAC. Similarly, the lead verb for 'V *into* n' is TAKE in both ICLE subsets but GO in LINDSEI. A concordance analysis shows that ICLE writers commonly use the phrase *take into account* as a discourse marker to summarize main points of their essays (e.g. *if we take into account all that has been said...*, ICLE German), while LINDSEI speakers use variations of GO *into* n, for example when talking about future plans in the warm-up section of the interviews (e.g. *I wanted to go into the pharmaceutical industry*, LINDSEI German).

Table 1: Type and token information for all VACs across ICLE and LINDSEI data sets (Ger = L1 German learners, Spa = L1 Spanish learners)

	ICLE_Ger			ICLE_Spa			LINDSEI_Ger			LINDSEI_Spa		
VAC	Types	Tokens	TTR	Types	Tokens	TTR	Types	Tokens	TTR	Types	Tokens	TTR
V *about* n	48	**242**	19.8%	42	**178**	23.6%	22	**147**	15.0%	17	**94**	18.1%
V *across* n	6	**9**	66.7%	0	**0**	0.0%	3	**4**	75.0%	0	**0**	0.0%
V *after* n	4	**15**	26.7%	1	**3**	33.3%	4	**5**	80.0%	2	**2**	100.0%
V *against* n	24	**45**	53.3%	13	**61**	21.3%	0	**0**	0.0%	1	**2**	50.0%
V *among* n	4	**5**	80.0%	5	**5**	100.0%	0	**0**	0.0%	0	**0**	0.0%
V *around* n	10	**14**	71.4%	10	**15**	66.7%	9	**18**	50.0%	4	**4**	100.0%
V *as* n	30	**56**	53.6%	30	**100**	30.0%	10	**15**	66.7%	8	**22**	36.4%
V *between* n	14	**22**	63.6%	11	**19**	57.9%	3	**3**	100.0%	1	**1**	100.0%
V *for* n	91	**338**	26.9%	78	**258**	30.2%	33	**98**	33.7%	20	**52**	38.5%
V *in* n	165	**556**	29.7%	163	**647**	25.2%	55	**344**	16.0%	38	**256**	14.8%
V *into* n	62	**175**	35.4%	25	**55**	45.5%	11	**33**	33.3%	3	**6**	50.0%
V *like* n	1	**1**	100.0%	2	**2**	100.0%	8	**57**	14.0%	5	**49**	10.2%
V *of* n	35	**149**	23.5%	44	**100**	44.0%	4	**16**	25.0%	4	**10**	40.0%
V *off* n	19	**32**	59.4%	4	**4**	100.0%	5	**5**	100.0%	1	**3**	33.3%
V *over* n	29	**45**	64.4%	6	**6**	100.0%	4	**6**	66.7%	0	**0**	0.0%
V *through* n	26	**40**	65.0%	15	**20**	75.0%	5	**7**	71.4%	4	**5**	80.0%
V *towards* n	11	**14**	78.6%	3	**3**	100.0%	0	**0**	0.0%	0	**0**	0.0%
V *under* n	11	**14**	78.6%	9	**18**	50.0%	1	**2**	50.0%	0	**0**	0.0%
V *with* n	111	**307**	36.2%	97	**269**	36.1%	26	**108**	24.1%	27	**117**	23.1%
Sum of hits		2079			1763			868			623	

Table 2: Overview of most frequent verb in each VAC across learner corpus datasets

VAC	**ICLE_Ger**	**LINDSEI_Ger**	**ICLE_Spa**	**LINDSEI_Spa**
V *about* n	THINK	TALK	TALK	TALK
V *across* n	(COME)	BE	n/a	n/a
V *after* n	LOOK	(BE)	(LOOK)	n/a
V *against* n	BE	(BE)	(FIGHT)	(GO)
V *among* n	(BE)	n/a	n/a	n/a
V *around* n	BE	(LOOK)	(BE)	n/a
V *as* n	WORK	(WORK)	APPEAR	WORK
V *between* n	DISTINGUISH	n/a	CHOOSE	n/a
V *for* n	WAIT	WORK	LOOK	BE
V *in* n	BE	BE	BE	BE
V *into* n	TAKE	GO	TAKE	(GO)
V *like* n	n/a	LOOK	n/a	BE
V *of* n	THINK	THINK	THINK	THINK
V *off* n	(FALL)	n/a	n/a	n/a
V *over* n	TAKE	(GO)	n/a	n/a
V *through* n	STROLL	(DRIVE)	n/a	(GO)
V *towards* n	(HEAD)	n/a	n/a	n/a
V *under* n	(BE)	(BE)	BE	n/a
V *with* n	DEAL	WORK	DEAL	BE

In an attempt to go beyond the single most frequent verb per VAC and gain further insights into learners' verb-VAC associations, we will now look at the top ten verb types used in the 'V *about* n' (displayed in Table 3) and the 'V *with* n' construction (Table 4) across learner corpora. Table 3 indicates overlap of the verbs most frequently found to occur in the 'V *about* n' VAC. THINK and TALK are the two most frequent verb lemmas in all four data sets, suggesting that, regardless of context or learner L1, these verbs of cognition and communication are strongly associated with this construction. With regard to the lower frequency verbs in the top ten lists, however, we observe corpus-specific variation. BE and SAY are more common in this VAC in the two LINDSEI datasets than in ICLE. In the spoken learner corpus, 'BE *about* n' is for example used to introduce topics of movies that the interviewees have seen (e.g. *it's a Mexican film*

and it's about two boys, LINDSEI Spanish). On the other hand, cognition verbs such as FORGET and HEAR are more commonly used in this VAC in ICLE than in LINDSEI. Both German and Spanish ICLE contributors use these verbs to make argumentative statements (e.g. *people waste their time and ... forget about all that really matters*, LINDSEI German) and to introduce a controversial issue they are writing about (e.g. *often we read and hear about this problem*, ICLE Spanish). Overall, corpus type or register seems to have a stronger effect on the verb-VAC selection than learner L1. This observation underscores the value of consulting different types of data in this kind of analysis. For the 'V *about* n' construction, either corpus would have been sufficient to uncover the most frequently selected verbs (THINK and TALK), but learners' register-specific verb-VAC associations may have been missed.

Table 3: Top ten verb choices for 'V *about* n' across learner corpus datasets

Rank	**ICLE_Ger**		**LINDSEI_Ger**		**ICLE_Spa**		**LINDSEI_Spa**	
1	THINK	67	TALK	40	TALK	49	TALK	23
2	TALK	36	THINK	38	THINK	37	THINK	20
3	CARE	17	BE	26	CARE	8	BE	14
4	FORGET	13	COMPLAIN	8	BRING	7	SPEAK	9
5	COMPLAIN	12	KNOW	6	SPEAK	7	COMPLAIN	5
6	KNOW	8	WORRY	6	WORRY	7	ARGUE	3
7	LEARN	7	LIKE	3	FORGET	6	SAY	3
8	BRING	6	SAY	3	KNOW	6	WORRY	3
9	HEAR	6	CARE	2	HEAR	5	CHOOSE	2
10	BE	5	LAUGH	2	BE	4	HEAR	2

The ICLE and LINDSEI results for 'V *with* n' present a similar picture of overlap of lead verbs and register-specific differences among other repeatedly used verbs, but also indicate differences between L1 German and L1 Spanish learner associations for this VAC. DEAL and LIVE are common associations across data sets. AGREE *with* n occurs more often in written than spoken learner data, which may not be surprising, given that ICLE is made up of argumentative essays in which learners state their opinions on controversial issues (e.g. *the main reason why I do not agree with military service…*, ICLE Spanish). The communication verb TALK, on the other hand, occurs more frequently in this VAC in both LINDSEI datasets than in ICLE. Verbs that differ across learner groups include COPE

and WORK, which are more often produced by German than Spanish learners, and the general, high-frequency verbs BE and GO which are the top two verbs in the LINDSEI Spanish list but are much less common in German learner production data. Another interesting verb in the ICLE Spanish list that does not appear in the German learner corpora is MARRY, as used in *she wants to marry with Hastings* (ICLE Spanish). This is likely an effect of crosslinguistic transfer from the learners' first language (Spanish *casarse con*, 'to marry with').

Table 4: Top ten verb choices for 'V *with* n' across learner corpus datasets

Rank	**ICLE_Ger**		**LINDSEI_Ger**		**ICLE_Spa**		**LINDSEI_Spa**	
1	DEAL	25	WORK	19	DEAL	31	BE	22
2	DO	20	LIVE	10	PLAY	25	GO	14
3	AGREE	19	DEAL	10	AGREE	22	LIVE	12
4	COPE	19	TALK	9	FINISH	12	DO	10
5	PLAY	13	STAY	7	HAPPEN	11	TALK	10
6	LIVE	10	START	6	WORK	8	WORK	7
7	WORK	9	GO	6	BEGIN	6	STAY	5
8	ASSOCIATE	8	COPE	6	MARRY	6	AGREE	4
9	COMMUNICATE	8	COME	5	BEHAVE	5	COMMUNICATE	4
10	GO	8	BE	5	DO	6	CONTINUE	4

4 Results (II): Experimental evidence on second language learner VAC knowledge

For additional evidence on L1 German and L1 Spanish learner knowledge of verbs in common VACs, we will now turn to the results from our lexical production task surveys. As in the previous section, we will first give an overview of verb frequencies for all VACs and list the number one verb per construction and participant group. We will then discuss survey participants' verb associations for two selected VACs (again 'V *about* n' and 'V *with* n') in a little more detail and comment on differences and similarities across groups.

Overall numbers of the verb types and tokens generated by survey participants in response to each VAC frame are shown in Table 5. Token numbers range from 290 for 'V *of* n' (an average production of three tokens per participant) in the L1 Spanish survey to 1,088 for 'V *across* n' in the native speaker survey (11 tokens on average). Some VACs triggered particularly high numbers

of responses from learners (especially 'V *in* n' and 'V *like* n'), while participants evidently found it harder to generate multiple verbs that they associate with other VACs, including 'V *of* n' and 'V *among* n'. If we compare the sums of token numbers across participant groups, we notice that, perhaps not surprisingly, both learner groups produced significantly lower numbers of verbs than the native speakers who completed the same survey and generated higher numbers of verbs in the 60 seconds available. It also appears that the L1 German learners found it slightly easier to think of a larger number of verbs than the L1 Spanish learners. Overall, token counts are much higher (and less varied) in the survey VAC results than in the learner corpora datasets, where numbers ranged from zero to 647.

Table 6 lists, for each VAC, the verb that each group of survey participants (native speakers, L1 German learners, L1 Spanish learners) generated most frequently when presented with a bare VAC frame. It can be argued that the top verb in terms of frequency rank is also the one that speakers (considered collectively) most strongly associate with a construction, given that it was generated by more participants than any of the other verbs. As we can see, there is considerable overlap between the German and Spanish survey results, with nine out of nineteen VACs sharing the same lead verb across L1 learner groups (e.g. LOOK for 'V *after* n', and GO for 'V *through* n'). In the L1 Spanish results, the dominant verbs across VACs are the general, high-frequency, semantically bleached verbs GO and BE (lead verbs in nine and seven VACs, respectively). Compared to their Spanish peers, German learners produce a more varied set of verbs that express more specific meanings of (directed) motion, including WALK, JUMP, and RUN. These lead verbs indicate a higher degree of overlap with the verbs most commonly generated by American English native speakers (see left-hand column in Table 6).

Table 5: Type and token information for all VACs across survey participant groups (NS = native speakers, Ger = L1 German learners, Spa = L1 Spanish learners)

	Survey_NS			Survey_Ger			Survey_Spa		
VAC	**Types**	**Tokens**	**TTR**	**Types**	**Tokens**	**TTR**	**Types**	**Tokens**	**TTR**
V *about* n	255	**740**	34.5%	144	**454**	31.7%	142	**436**	32.6%
V *across* n	277	**1088**	25.5%	161	**582**	27.7%	105	**478**	22.0%
V *after* n	320	**950**	33.7%	166	**529**	31.4%	146	**506**	28.9%
V *against* n	243	**779**	31.2%	159	**512**	31.1%	156	**473**	33.0%
V *among* n	303	**866**	35.0%	142	**410**	34.6%	154	**445**	34.6%
V *around* n	317	**999**	31.7%	154	**559**	27.5%	155	**536**	28.9%
V *as* n	358	**879**	40.7%	215	**479**	44.9%	187	**521**	35.9%
V *between* n	312	**937**	33.3%	183	**555**	33.0%	153	**511**	29.9%
V *for* n	274	**866**	31.6%	192	**520**	36.9%	156	**464**	33.6%
V *in* n	318	**990**	32.1%	217	**617**	35.2%	183	**584**	31.3%
V *into* n	290	**996**	29.1%	168	**615**	27.3%	135	**537**	25.1%
V *like* n	275	**891**	30.9%	151	**651**	23.2%	185	**668**	27.7%
V *of* n	166	**435**	38.2%	119	**293**	40.6%	147	**290**	50.7%
V *off* n	283	**912**	31.0%	131	**475**	27.6%	113	**481**	23.5%
V *over* n	318	**1024**	31.1%	174	**577**	30.2%	157	**594**	26.4%
V *through* n	317	**1078**	29.4%	181	**585**	30.9%	138	**508**	27.2%
V *towards* n	273	**971**	28.1%	163	**545**	29.9%	122	**439**	27.8%
V *under* n	306	**995**	30.8%	174	**587**	29.6%	154	**533**	28.9%
V *with* n	342	**997**	34.3%	216	**591**	36.5%	204	**590**	34.6%
Sum of responses		17393			10136			9594	

Table 6: Overview of most frequent verb in each VAC across survey participant groups

VAC	**Survey_NS**	**Survey_Ger**	**Survey_Spa**
V *about* n	TALK	TALK	TALK
V *across* n	RUN	WALK	GO
V *after* n	RUN	LOOK	LOOK
V *against* n	FIGHT	BE	BE
V *among* n	BE	BE	BE
V *around* n	RUN	WALK	GO
V *as* n	RUN	LOOK	BE
V *between* n	RUN	BE	BE
V *for* n	RUN	LOOK	GO
V *in* n	RUN	BE	BE
V *into* n	RUN	RUN	GO

V *like* n	LOOK	LOOK	LOOK
V *of* n	BE	THINK	BE
V *off* n	FALL	FALL	GO
V *over* n	JUMP	JUMP	GO
V *through* n	RUN	GO	GO
V *towards* n	RUN	RUN	GO
V *under* n	RUN	BE	BE
V *with* n	RUN	GO	GO

If we now compare the ten verbs most frequently generated by L1 German and L1 Spanish learners in response to the 'V *about* n' prompt, we notice a lot of overlap among the verbs with the highest token numbers. Table 7 lists the ten most frequent verb responses for this VAC across learner groups, with the native speakers' responses provided for reference purposes (our focus is not on comparing learners against native speakers, but on what the survey data tell us about learners' verb-VAC associations). Seven of the top ten verbs associated with this VAC are shared among the two learner groups. Five of these (TALK, BE, THINK, SPEAK, and WRITE) are also the verbs with the highest absolute token frequencies and hence the most entrenched items for the 'V *about* n' pattern. Both learner groups associate with this VAC verbs of communication and cognition. Associations between the VAC and TALK, SPEAK, and THINK seem to be particularly strong. Two of the most common verbs in the native speaker responses, RUN and WALK, are rare among the learner responses (with between two and five instances). The directed motion sense expressed by these verbs does not appear to be a sense that this VAC activates in the minds of the learners who participated in the survey. Also interesting is the verb DISCUSS in the L1 Spanish top ten list (generated by nine participants). As we observed elsewhere (Römer, O'Donnell and Ellis forthcoming), a prescriptive grammar would consider this use ungrammatical, but we may in fact be witnessing the development of a new phrasal verb – especially given additional attestations of DISCUSS *about* in ESL and native English varieties.

Table 7: Top ten verb responses for 'V *about* n' across survey participant groups

Rank	**Survey_NS**		**Survey_Ger**		**Survey_Spa**	
1	TALK	40	TALK	55	TALK	53
2	RUN	35	BE	44	BE	46
3	THINK	35	THINK	32	THINK	30
4	WALK	27	SPEAK	19	SPEAK	24
5	BE	26	WRITE	16	COME	12
6	SPEAK	26	SING	13	WRITE	12
7	GO	20	READ	12	GO	10
8	WONDER	17	KNOW	8	ASK	10
9	WRITE	15	ASK	8	DISCUSS	9
10	CRY	13	LAUGH	8	KNOW	7

Table 8 compares the ten most frequent verb responses for the 'V *with* n' VAC across participant groups. As in the case of 'V *about* n', German and Spanish learners' verb-VAC associations overlap considerably. Eight of the top ten verbs are shared across the two groups. In addition to verbs which express general directed motions (RUN, GO, WALK) or group activities (PLAY, WORK) and which are also frequently generated by native speaker survey participants, both learner groups often produce forms of the verbs COME and BE. The verb SING, which is unique to the L1 German top ten list, may be the result of cross-linguistic transfer (German *mitsingen*, 'to sing with'). All verbs produced by the two learner groups share high frequencies in general English usage.

Table 8: Top ten verb responses for 'V *with* n' across survey participant groups

Rank	**Survey_NS**		**Survey_Ger**		**Survey_Spa**	
1	RUN	47	GO	41	GO	50
2	WALK	35	COME	32	COME	35
3	GO	33	RUN	25	BE	30
4	EAT	23	BE	22	LIVE	20
5	PLAY	22	PLAY	22	PLAY	20
6	TALK	21	TALK	20	RUN	19
7	SLEEP	19	WALK	19	TALK	17
8	JUMP	19	WORK	16	WALK	16
9	WORK	18	MOVE	15	WORK	16
10	FLY	17	SING	14	EAT	13

5 Comparison of results and evaluation of data types

If we now compare the findings on learner VAC knowledge obtained in the learner corpus analysis with those gathered in the lexical production tasks, we notice both similarities and differences in verb selection preferences. Starting with a comparison of the overall lead verb patterns in survey and corpus data (see Tables 2 and 6), we find an overlap in types for most VACs for which there were sufficient attestations in the learner corpora. The lead verb for 'V *about* n' across data sets is TALK, the one for 'V *in* n' is BE, and GO for 'V *into* n'. For these constructions, either experimental or corpus evidence would have been sufficient to identify the most commonly associated verb. For the commonly attested VAC 'V *with* n', however, the number one verb is different across data sets and appears to depend on the source of data and the task reflected therein (DEAL in ICLE, WORK/BE in LINDSEI, GO in survey data). Low token numbers for the majority of selected VACs in the two learner corpora preclude a more comprehensive comparison of lead verbs.

The analysis of top ten verb lists for 'V *about* n' and 'V *with* n' across data types (compare Tables 3 and 7, as well as Tables 4 and 8) enable us to look beyond the lead verb and indicate further differences between spoken and written learner production on the one hand and survey responses on the other. While TALK and THINK occur at the top of the L1 German and L1 Spanish lists across data sets for 'V *about* n', we find that other verbs in this VAC depend on the types of text captured in ICLE and LINDSEI (e.g. COMPLAIN, ARGUE, WORRY). The survey data do not highlight these text-type or task-specific verbs but instead provide us with an additional set of communication verbs (SPEAK, WRITE, READ, ASK) – verbs which are semantically related to the lead verb TALK. Learner corpus and survey results also differ when it comes to learners' more specific verb-VAC associations for 'V *with* n'. We already commented on the lack of overlap across lead verbs in this VAC and mentioned that its use appears to be register-dependent. One common verb that occurs in top ten lists across data types is WORK. Both learner corpus and survey data indicate that WORK *with* n is a pattern that is easily retrievable for learners. Other verbs that occur repeatedly in this VAC differ across data types. The survey results suggest that learners (of both L1s) strongly associated motion verbs such as RUN, WALK, and MOVE with this VAC, but these verbs are not at all common in data sets retrieved from the learner corpora where verbs such as DEAL, COPE, and LIVE (ICLE/LINDSEI German) and AGREE and DO (ICLE/LINDSEI Spanish) form patterns with this VAC. This again shows an effect of text type or task performed by the learners on the results.

This comparison of results from different data sources highlights a few important issues related to working with small learner corpora in studying learner knowledge of VACs and underscores, in our opinion, the usefulness of a combined 'corpus plus experimental data' approach. One major difference between the two data types we used became apparent in the overviews of verb frequencies presented at the beginning of the two results sections (Tables 1 and 5). While it is often the case that corpora afford the researcher "a larger range of data […] than many experimental designs allow for" (Gilquin and Gries 2009: 8-9), we have found the opposite to be true in this study. ICLE and LINDSEI yielded fairly small token and type numbers for most VACs, relative to experimental data. Since verbs in VACs display a Zipfian pattern of distribution with a few verb types accounting for the vast majority of tokens and many other verb types appearing only once or twice (Ellis, O'Donnell and Römer 2013), larger token numbers are essential if we wish to extend our analysis beyond the lead verb type and identify sets of semantically related verbs.

A potential downside of relying entirely on survey data is that its level of naturalness is lower than that of learner corpus data. Participants supply lists of verbs in a bare, decontextualized frame under time pressure, rather than performing a task they are used to, such as writing about their opinions on a certain topic (ICLE) or talking to someone about their experiences or things they like (LINDSEI). As our analysis highlighted, however, the fact that language in ICLE and LINDSEI is contextualized can also be considered a drawback. For several of the selected VACs we found a register or task effect at work, so that verbs which repeatedly occurred in a VAC do so to form phrases that fulfill a register-specific communicative purpose. The more controlled survey setting, on the other hand, arguably produces more 'neutral' sets of verbs for each VAC that mirror more accurately what learners' lexical associations with those constructions look like.

6 Conclusion and outlook

This paper has explored the value of combining learner corpus and experimental evidence in the context of studying L2 learners' knowledge of a selection of commonly used English verb-argument constructions (VACs). Analyses of VACs in the German and Spanish components of ICLE and LINDSEI and of the responses of L1 German and L1 Spanish learners to a lexical production task have helped us identify which verbs advanced L2 speakers of English most strongly associate with particular constructions.

Both types of data provide evidence for the entrenchment of VACs in the learners' minds (supporting findings discussed in Ellis, O'Donnell and Römer submitted, and in Römer, O'Donnell and Ellis forthcoming). Both groups of learners have constructional knowledge, use a small set of verbs repeatedly, and show associations of VACs with verbs that belong to VAC-specific semantic groups (e.g. verbs of cognition or communication for 'V *about* n'). The most common verbs tend to be shared across data sets, which means that it is possible to identify the most strongly associated verbs in a VAC through either a learner corpus or a survey data analysis. However, if we had relied exclusively on the ICLE/LINDSEI analyses, we would have missed a number of verbs that learners strongly associate with common VACs (verbs that they generated repeatedly in the lexical production task) and which contribute to the meaning of these VACs, and hence to our understanding of what learners know about them. The main reason for this lies in the scarcity of occurrences of a large number of VACs in the learner corpora we had access to. The psycholinguistic experiments helped us gather larger data sets which enabled us to gain additional insights into how entrenched selected VACs are in the second language learner's mind and which verbs learners most strongly associate with particular VACs. The larger survey data sets covered more verb types and tokens per VAC than those based on the learner corpora. Semantic analyses for which the ICLE and LINDSEI data sets are simply too small can be carried out based on the survey results.

As Gilquin and Gries (2009: 9) aptly noted, "like any other method in isolation, corpora are not perfect." We certainly found this to be true in our study. We join other researchers who have made a case for combining different types of evidence on a linguistic phenomenon and have commented on the benefits of mixed- or multi-methods research (Wulff 2009; Arppe et al. 2010; Gries 2013). We recognize the compatibility of corpus and cognitive approaches to language analysis and believe that combinations of corpus and experimental evidence will lead to richer results and enable us to enhance work in Corpus Linguistics.

We are aware that there are limitations to our study and that, in order to fully understand what learners know about English verb-argument constructions, more research is required. A logical next step for us is to include additional VACs in our analysis. We have already collected survey data on an additional set of 20 VACs and are now mining ICLE and LINDSEI for those same constructions. We are looking to include other, larger corpora that contain data produced by German and/or Spanish learners of English as well. There is currently a lack of freely available large (and particularly of dense) corpora of learner English, so we are hoping to see more development in this area of resource creation and sharing. Also, the data sets resulting from the verbal flu-

ency tests are much richer than this paper suggests. Our next analytic steps will therefore include a more detailed analysis of those data, which considers the order of responses, participants' reaction times, systematic comparisons of native speaker and learner responses, and VAC-specific groupings of verbs into semantic sets. Differences across native speaker and learner responses need to be statistically quantified, which will include calculating and plotting correlations between data sets (as done in Römer, O'Donnell and Ellis forthcoming, submitted). So, our explorations of speaker knowledge of VACs continue… Meanwhile, we hope that we have inspired others to carry out research which combines different types of linguistic evidence so that we will be able to witness more of the synergetic effects of data triangulation which we believe will lead to a further maturation of the field of Corpus Linguistics.

Note

* The authors would like to thank Ben Pinkasovic for his invaluable help with the administration of the VAC knowledge surveys and the recruitment of participants. We would also like to thank the following friends and colleagues for their help with distributing VAC surveys to students in Germany, Spain, and the United States: Carmen Aguilera Carnerero, Rafael Alejo Gonzalez, Ulrike Altendorf, Laura Aull, Ruth Breeze, Scott Crossley, Belén Diez-Bedmar, Fiorella Dotti, Izis Elorza, Encarna Hidalgo, Lars Hinrichs, Matt Jadlocki, Sarah Kegley, Daniela Kolbe-Hanna, Rolf Kreyer, Joseph Lee, María José López-Couso, Isabel Miñés, Juan Carlos Palmer, Caroline Payant, Carmen Perez-Llantada, Pascual Perez Paredes, Miguel Ruiz Garrido, Carmen Sancho Guinda, Andrea Sand, Marco Schilk, Rainer Schulze, Ayush Shrestha, Mary Smith, and Stefanie Wulff. The lead author acknowledges the support of the project 'Measuring speakers' knowledge of English verb-argument constructions: Psycholinguistic evidence from first and second language settings' through the Georgia State University C. F. Arrington Research Initiation Grant Program.

References

Arppe, Antti, Gaëtanelle Gilquin, Dylan Glynn, Martin Hilpert and Arne Zeschel. 2010. Cognitive corpus linguistics: Five points of debate on current theory and methodology. *Corpora* 5 (1): 1–27.

Bird, Steven, Edward Loper and Ewan Klein. 2009. *Natural language processing with Python*. Cambridge, MA: O'Reilly Media Inc.

Ellis, Nick C., Matthew B. O'Donnell and Ute Römer. 2013. Usage-based language: Investigating the latent structures that underpin acquisition. *Language Learning* 63 (Supp. 1): 25–51.

Ellis, Nick C., Matthew B. O'Donnell and Ute Römer. Forthcoming. The processing of verb-argument constructions is sensitive to form, function, frequency, contingency, and prototypicality. *Cognitive Linguistics*.

Ellis, Nick C., Matthew B. O'Donnell and Ute Römer. Submitted. Second language processing of verb-argument constructions is sensitive to form, function, frequency, contingency, and prototypicality.

Ellis, Nick C. and Rita Simpson-Vlach. 2009. Formulaic language in native speakers: Triangulating psycholinguistics, corpus linguistics, and education. *Corpus Linguistics and Linguistic Theory* 5 (1): 61–78.

Francis, Gill, Susan Hunston and Elizabeth Manning. 1996. *Grammar patterns 1: Verbs*. London: Harper Collins.

Gilquin, Gaëtanelle and Stefan Th. Gries. 2009. Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory* 5 (1): 1–26.

Gilquin, Gaëtanelle, Sylvie De Cock and Sylviane Granger (eds.). 2010. *LINDSEI: Louvain International Database of Spoken English Interlanguage*. Louvain-la-Neuve, Belgium: Presses Universitaires de Louvain.

Granger, Sylviane, Estelle Dagneaux, Fanny Meunier and Magali Paquot (eds.). 2009. *ICLE: International Corpus of Learner English*. Louvain-la-Neuve, Belgium: Presses Universitaires de Louvain.

Gries, Stefan Th. 2013. Data in Construction Grammar. In G. Trousdale and T. Hoffmann (eds.). *The Oxford handbook of Construction Grammar*, 93–108. Oxford: Oxford University Press.

Gruenewald, Paul J. and Gregory R. Lockhead. 1980. The free recall of category examples. *Journal of Experimental Psychology: Human Learning and Memory* 6 (3): 225–240.

Römer, Ute, Matthew B. O'Donnell and Nick C. Ellis. 2012. What do speakers know about English Verb-Argument constructions? Combining corpus and psycholinguistic evidence from L1 and L2 settings. Paper presented at ICAME 33 in Leuven, Belgium, May 2012.

Römer, Ute, Matthew B. O'Donnell and Nick C. Ellis. Forthcoming. Using COBUILD grammar patterns for a large-scale analysis of verb-argument constructions: Exploring corpus data and speaker knowledge. In M. Charles, N. Groom and S. John (eds.). *Corpora, grammar, text and discourse: In honour of Susan Hunston.* Amsterdam: John Benjamins.

Römer, Ute, Matthew B. O'Donnell and Nick C. Ellis. Submitted. Second language learner knowledge of verb-argument constructions: Effects of language transfer and typology.

Wulff, Stefanie. 2009. Converging evidence from corpus and experimental data to capture idiomaticity. *Corpus Linguistics and Linguistic Theory* 5 (1): 131–159.