Ute Römer*, Stephen C. Skalicky and Nick C. Ellis

Verb-argument constructions in advanced L2 English learner production: Insights from corpora and verbal fluency tasks

https://doi.org/10.1515/cllt-2016-0055

Abstract: This paper draws on data from learner and native-speaker corpora as well as psycholinguistic data to gain insights into second language speaker knowledge of English verb-argument constructions (VACs). For each of 34 VACs, L1 German and L1 Spanish advanced English learners' and English native speakers' dominant verb–VAC associations are examined based on data retrieved from the International Corpus of Learner English (ICLE), the Louvain International Database of Spoken English Interlanguage (LINDSEI), their respective Native Speaker (NS) reference corpora, and data collected in verbal fluency tasks in which participants complete VAC frames, such as, 'she _____ with the...' with verbs that come to mind. We compare findings from the different data sets and consider the strengths and limitations of each in relation to questions in usage-based second language acquisition and Construction Grammar.

Keywords: construction grammar, verb-argument constructions, second language acquisition, learner corpora, psycholinguistic evidence

1 Introduction

There has been a recent increase in studies that highlight the value of combining corpus and experimental evidence in the study of linguistic phenomena (e.g., Arppe et al. 2010; Ellis and Simpson-Vlach 2009; Ellis et al. 2016; Gilquin 2007; Gilquin and Gries 2009; Gries and Wulff 2005, 2009; Gries et al. 2005, 2010; Mollin 2014; Rebuschat et al. 2017; Wulff 2008, 2009). These studies utilize corpora, such as the British National Corpus (BNC) or the British component of the International Corpus of English (ICE-GB), in combination with speaker

***Corresponding author: Ute Römer,** Department of Applied Linguistics and ESL, Georgia State University, Atlanta, GA, USA, E-mail: uroemer@gsu.edu
Stephen C. Skalicky, Department of Applied Linguistics and ESL, Georgia State University, Atlanta, GA, USA, E-mail: sskalicky1@gsu.edu
Nick C. Ellis, Departments of Psychology and Linguistics, University of Michigan, Ann Arbor, MI, USA, E-mail: ncellis@umich.edu

judgments collected in experimental settings and demonstrate that different types of data present converging evidence to strengthen research hypotheses. These studies also show that combining data types enables us to ask questions that one data type alone is not sufficient to address. With few recent exceptions (e.g., Römer et al. 2014b; Littré 2015), the focus in such studies has been on corpora that capture *native speaker output*. The present paper discusses how corpora which capture *second language learner output* can be used in combination with psycholinguistic data to obtain better insight into second language acquisition.

Our goal is to draw on corpus as well as psycholinguistic evidence in order to provide insights into second language (L2) learner knowledge of 34 different English verb-argument constructions (henceforth VACs), for example the 'V *with* n' construction (as exemplified by *she dealt with the issue*) and the 'V reflexive pronoun' construction (exemplified by *he contradicts himself constantly*). Verb constructions are central building blocks of language, and associations between verbs and constructions have been shown to differ across L1 and L2 speaker groups, with learners of L1s that differ typologically from the L2 struggling more to produce target-like verbs than learners whose L1 is typologically related to the target language (Ellis et al. 2014b). Crosslinguistic influence is of key importance in second language development as it may either support or hinder the acquisition process (Gass and Selinker 1983; Jarvis 2011, 2013; Jarvis and Pavlenko 2008; Odlin 2013). An L2 learner's mind is not a *tabula rasa*. The learner's knowledge of constructions in their first language is likely to have an impact on their emerging constructional knowledge in the L2. Also, research in usage-based SLA has shown that each language leads its speakers to experience different "thinking for speaking" (Cadierno 2013; Slobin 1996) and hence to construe the world in different ways. When a speaker learns another language, this process involves learning how to construe the world like native speakers and, along the way, learning an alternate way of (re-)thinking for speaking (Brown and Gullberg 2008, 2010; Cadierno 2008; Robinson and Ellis 2008).

For these reasons, we consider it important to learn more about preferred verb-VAC combinations in L2 learner production data and about how those differ from the verb-VAC associations of native speakers. This may contribute to a better understanding of second language acquisition. Our research questions are:

RQ1: Which verbs do advanced L1 German and L1 Spanish learners of English most commonly associate with a particular VAC?

RQ2: How do learners' verb-VAC associations compare to those of native speakers? and

RQ3: How well do learner corpus and psycholinguistic data complement each other in providing insights into L2 learner VAC knowledge?

Finding answers to these questions will be important as we work toward a broader range of mixed-methods approaches in corpus linguistics which combine various types of empirical evidence, while also allowing us to gain better insight into L2 learners' VAC knowledge.

The context of our study is a collaborative project which investigates the use, processing, and acquisition of VACs (described in detail in Ellis et al. 2016). The project takes constructions identified and discussed in the *COBUILD Grammar Patterns: Verbs* volume (Francis et al. 1996) as a starting point for a systematic analysis of VACs in the BNC (see Römer et al. 2015 for a description of the BNC VAC extraction). By means of free-association data, the project also studies native-speaker and learner associations of verbs and VACs. Comparisons of the results from the BNC analyses and the free-association experiments allow us to determine in what ways and to what extent speakers are affected by verb distributions in the input (see Ellis et al. 2013, 2014a). Central findings of the study include: (1) verb-argument constructions are Zipfian in their type-token distribution, with one verb type accounting for the lion's share of all VAC tokens, (2) they are selective in their verb form occupancy, and (3) they are coherent in their semantics (Ellis et al. 2013). Subsequent comparisons of BNC frequencies against the same free-association 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 et al. 2014a, 2014b).

To investigate what L2 learners of English know about VACs and the verbs that tend to appear in them, our previous work collected and evaluated data from free association tasks in which L2 English learners generate the first word that came to mind to fill the verb slot in 20 sparse VAC frames such as 'She _____ about the...' (see Römer et al. 2014b). The study we report on in the present paper goes beyond this work by including a richer type of psycholinguistic data (verbal fluency tasks) and additional data on VACs retrieved from spoken and written learner corpora. Selected results from a pilot study based on those two new data types have been reported in Römer et al. (2014b). While this pilot study served to partly address RQ1 and RQ3 above, it did so unsatisfactorily, with a small-scale scope, a focus on qualitative analysis, and with a lack of quantitative evaluation of the data. It also did not consider native-speaker reference data to compare the learner production data against, and hence did not allow insights on differences or similarities between L1 and L2 English speakers' verb-VAC associations (our RQ2). Previous publications by our research team included such comparisons, but only based on data from one

type of free association task (Ellis et al. 2014b, 2016; Römer et al. 2014a). Richer psycholinguistic data or learner corpus data were not considered. In the comparison of L1 and L2 English speakers' verb-VAC associations, our previous work also only covered 19 constructions or the 'V preposition n' type. The study described in the current paper goes beyond our previous work in the following ways: it (1) draws upon data for 34 VACs of different types, (2) collects speaker production evidence on those VACs from lexical fluency tasks, spoken and written learner corpora, and spoken and written native-speaker corpora designed to match the learner corpora, and (3) employs more systematic quantitative methods of data processing, evaluation, and visualization. This study hence allows us to address our research questions more fully and provide more robust evidence on VACs in L2 English learner production.

In what follows, we first describe our data types and analytic methods, then present central results from our analyses, discussing the usefulness of combining corpus and experimental data in our endeavor, and conclude with a summary of observations and thoughts on future research tasks on the topic.

2 Data and methods

In order to address our research questions and better understand learner knowledge of VACs, we analyzed two main types of data: (1) data extracted from spoken and written corpora and (2) data produced in free-association fluency experiments. Both data types (and subtypes as described in Sections 2.1 and 2.2) were analyzed quantitatively using type-token distributions and correlation analysis, as well as qualitatively by comparing frequency-ranked lists of verbs across datasets. Table 1 provides an overview of the data types and sources that our study is based on.

The two selected data types (i.e., corpus and psycholinguistic) differ in terms of naturalness, or in the extent to which they mirror authentic communicative practices. Corpora can be considered a much more natural source of linguistic data than psycholinguistic experiments, with prototypical native-speaker corpora such as the British National Corpus (BNC) or the Corpus of Contemporary American English (COCA) being more natural than less prototypical corpora such as learner corpora (see Gilquin and Gries 2009). While learner corpora consist of authentic texts, these texts have usually not been produced in natural communicative settings but instead for the purpose of language learning, practice, or assessment. Gilquin and Gries rank psycholinguistic experiments "requiring subjects to do something with language they usually do

Table 1: Data sources used in this study.

(a) Corpus data			
Corpus	Register type	Speaker/Writer L1	Corpus size
ICLE_German	Writing	German	236,095 words
ICLE_Spanish	Writing	Spanish	198,109 words
LOCNESS	Writing	English	264,095 words
LINDSEI_German	Speech	German	86,072 words
LINDSEI_Spanish	Speech	Spanish	63,889 words
LOCNEC	Speech	English	118,564 words
(b) Psycholinguistic data			
Task type	Subject L1	Number of subjects (1^{st} set of VACs)	Number of subjects (2^{nd} set of VACs)
Verbal fluency task	German	94	97
Verbal fluency task	Spanish	96	97
Verbal fluency task	English	94	99

not do" (2009: 5) among the least natural types of linguistic data, certainly much less natural than learner corpora. We will come back to the issue of naturalness in our discussion of the usefulness of selected data sources in the results section.

2.1 Corpus data extraction

We collected corpus data from four sources: ICLE, LINDSEI, LOCNESS, and LOCNEC. We first extracted VACs from subsections of the International Corpus of Learner English (ICLE) and the Louvain International Database of Spoken English Interlanguage (LINDSEI). ICLE consists of mostly argumentative essays written by advanced undergraduate EFL learners from 16 different L1 backgrounds (Granger et al. 2009). For our analysis, we selected the German and Spanish native speaker sections (henceforth ICLE_German and ICLE_Spanish), consisting of 437 and 251 texts respectively (comprising 236,095 and 198,109 words). To facilitate VAC extraction from ICLE_German and ICLE_Spanish, we part-of-speech (POS) tagged both corpora using TagAnt (Anthony 2014b).

Spoken learner production data came from LINDSEI, which contains transcripts of informal interviews with EFL undergraduate students from 11 L1 backgrounds. The majority of students' language proficiency in LINDSEI was rated as high intermediate (Gilquin et al. 2010). LINDSEI tasks include a free warm-up conversation on a topic selected by the learners (either a country they have visited, a lesson they have learned, or a film they have seen), followed by a

discussion about the chosen topic, followed by a picture description task. As with ICLE, we used the L1 German and L1 Spanish subsets of this corpus (henceforth LINDSEI_German and LINDSEI_Spanish), each consisting of 50 interviews, and only included speech produced by learners (86,072 and 63,889 words respectively). For LINDSEI_German and LINDSEI_Spanish we decided to work with the plain unannotated text files, given that automatic POS tagging of spoken learner language can be rather error-prone.

We also extracted VAC data from the native-speaker reference corpora to ICLE and LINDSEI: the Louvain Corpus of Native English Essays (LOCNESS, Granger 1996) and the Louvain Corpus of Native English Conversations (LOCNEC, De Cock 2004). Like ICLE, LOCNESS consists of student argumentative essays. We only included the LOCNESS subsets that contain British and American university students' essays (264,095 words overall), excluding essays written by British A-levels students. We POS-tagged LOCNESS using TagAnt. LOCNEC matches LINDSEI in its design and is made up of transcripts of 50 interviews with British native speakers of English (118,564 words). As with LINDSEI and for the same reason, we decided against automatically tagging LOCNEC.

The 34 VACs for which we retrieved examples from ICLE_German, ICLE_Spanish, LINDSEI_German, LINDSEI_Spanish, LOCNESS, and LOCNEC are listed in Table 2. All VACs were selected from the *COBUILD Grammar Patterns* volume on verbs (Francis et al. 1996). To extract candidate lists of

Table 2: The 34 VACs included in our analysis (in alphabetical order).

it is V-ed *that*	V *in* n
it V *like*	V *into* n
pl-n V *together*	V *like* n
there V n	V n *against* n
V *about* n	V n *amongst* n
V *across* n	V n *around* n
V *after* n	V *not*.
V *against* n	V *of* n
V *ahead*	V *off* n
V *among* n	V *over* n
V *around* n	V reflexive pronoun
V *as if*	V *so*.
V *as* n	V *through* n
V *between* n	V *toward(s)* n
V *down* n	V *under* n
V *for* n	V *way* prep
V *in favor of* n	V *with* n

VAC examples in context, we used the concordance tool AntConc (Anthony 2014a) and searched for lexical strings that would ensure maximum recall, for example *about, as, in favor of*, or *there*. In the case of ICLE_German, ICLE_Spanish, and LOCNESS, we were able to search for combinations of verb plus lexical item (*across, like, so*, etc.), for example "*_V* so_*". The resulting concordances then required manual filtering for true instances of each VAC, especially in the case of LINDSEI/LOCNEC where we were not able to carry out POS searches. For example, we excluded instances for 'V *about* n' in which a preposition was used as an adverb, as in *Theresa was about only twenty-five* (ICLE_German), or in which an element of the VAC was missing, as in *it's rarely talked about* (ICLE_German; missing noun/noun group after *about*). For all VACs and datasets we then created lemmatized, frequency-sorted verb lists (204 lists altogether; 34 times six corpora).

2.2 Psycholinguistic data

The second data type we draw upon in our study comes from verbal fluency tasks, which help investigate semantic category knowledge and its fluency of access. These tasks ask participants to generate as many words as come to mind within a given time (usually 60 seconds) when presented with a certain stimulus (usually a category label such as "animals", or "words that start with *gl*"). Results from verbal fluency tasks provide insights into the typicality of items in a certain context (or of a certain category), as more typical items are produced by more participants and earlier within the given time frame than less prototypical ones (Gruenewald and Lockhead 1980; Kail and Nippold 1984). We adopted this task to further probe speakers' VAC knowledge and see which verbs (or verb groups) learners and native speakers most commonly associate with the 34 VACs listed in Table 2.

In our verbal fluency task, we presented three groups of participants (native speakers, L1 German and L1 Spanish learners of English) with bare sentence frames based on 40 different VACs consisting of a singular pronoun subject (*he/she/it*), an empty slot for the verb, and the lexical elements of the respective VAC. Example VAC frames used in our survey were 'she _____ about the … ', 'it _____ itself', and 'he _____ it among the … '.[1] The tasks were counterbalanced to show both 'he/she' and 'it' variants at equal frequency. The survey was

1 Data on six of the 40 survey VACs that were not included in the corpus analysis is not covered in this paper. These six VACs (e.g., 'V n', 'V n n') were deemed to be particularly difficult to extract automatically from corpora with good precision and high recall.

designed and delivered online via Qualtrics (www.qualtrics.com). Participants in all three groups were college students mostly in their twenties, recruited via email from universities in the US, Germany and Spain. The Qualtrics survey links were only shared with native English speakers in the US and with advanced learners of English in Germany and Spain. The majority of learners were at proficiency level C1 in the Common European Framework of Reference for Languages (CEFR) except for a small number of Spanish learners (less than 10%) who were at CEFR level B2. To compensate the survey participants for their time, we offered participants an Amazon gift card (worth five USD or five EUR). All participants were instructed to type in the first verbs they could think of that could fill the gap in the VAC frames they saw and to press the enter key after each verb. They were informed that the 60 second countdown would start as soon as they began to type the first verb and that it was okay to take breaks between prompts.

The VACs were split into two groups of 20, and each participant completed verbal fluency tasks for 20 VACs in random order. Because each of the VAC subgroups had a different total numbers of participants (94 or 99 American English native speakers, 94 or 97 advanced L1 German speakers, and 96 or 97 advanced L1 Spanish speakers), not every VAC was displayed an identical number of times. For example, 97 German learners saw the prompt for 'V reflexive pronoun' but only 94 German learners saw the prompt for 'V *about* n'. For each trial, we recorded participants' verb responses and the time they took between verb responses for each VAC frame. We lemmatized the responses from each of the three subject groups using the Natural Language Toolkit (NLTK; Bird et al. 2009) and created verb frequency lists for each VAC and participant group (120 lists altogether; 40 VACs times three L1 groups).

2.3 Data analysis

The frequency-sorted verb type-token lists based on VAC data extracted from corpora and produced in the verbal fluency task allowed for a range of comparisons, including both quantitative and qualitative analyses. We compared verb type and token lists (1) across VACs, (2) across learner and native speaker datasets, (3) across L1 learner groups, and (4) across data types. For these comparisons and for VACs that produced sufficient numbers of hits in the corpora and/or responses in the verbal fluency tasks, we carried out a correlation analysis and visualized correlations by means of scatterplots using R (R Development Core Team 2012). For a small group of VACs, we also looked more qualitatively at the top verb choices across data sets to better understand

how learner verb-VAC associations compare to those of native speakers. In the following, we discuss selected representative results from our quantitative and qualitative analyses.

3 Results

We first present an overview of verb type and token distributions across all 34 VACs for datasets gathered from written corpora, spoken corpora, and verbal fluency tasks. We then zoom in on three VACs that are particularly frequently attested in our datasets: 'V *about* n', 'V *with* n', and 'V reflexive pronoun'.

3.1 Overview of VACs across datasets

The data extraction and collection procedures described in Sections 2.1 and 2.2 resulted in frequency-sorted verb lists (for each VAC and dataset) that we analyzed in terms of types and tokens. We did this separately for the written corpora (ICLE_German, ICLE_Spanish, LOCNESS), spoken corpora (LINDSEI_German, LINDSEI_Spanish, LOCNEC), and verbal fluency task data (Survey_German, Survey_Spanish, Survey_NS). The goals were to determine (1) how frequently each of the selected VACs occurred in learner production data (with NS production data collected for reference purposes), (2) how productive each VAC was in terms of verb types that occurred in it, and (3) what the most frequent verbs were in each VAC. We started with an analysis of the learner corpus verb lists and then looked for additional evidence of learner VAC knowledge in the psycholinguistic verbal fluency data.

Table 3 shows frequency information for the VACs we extracted from ICLE_German, ICLE_Spanish, and LOCNESS. For each VAC and corpus we list token numbers (i.e., how many instances of a VAC we identified), type numbers (i.e., how many different verbs were used in a set of VAC tokens), and calculated type-token ratios which indicate how productive a VAC is (but have to be treated with caution in the case of VACs with low token frequencies). What we see in Table 3 is that verb type and token numbers differ widely across VACs and range from 0 ('V *across* n' in ICLE_Spanish) to 997 tokens and 302 types ('V *in* n' in LOCNESS). In addition to 'V *in* n', we found high numbers of instances in the written corpora for '*there* V n', 'V *for* n', 'V *into* n' (although less frequent in ICLE_Spanish), 'V *of* n', 'V reflexive pronoun', and 'V *with* n'. All other VACs are fairly infrequent in learner and NS writing. This especially applies to 'V *ahead*',

Table 3: Verb type and token frequencies across VACs in written corpora.

	ICLE_German			ICLE_Spanish			LOCNESS		
VAC	**Types**	**Tokens**	**TTR**	**Types**	**Tokens**	**TTR**	**Types**	**Tokens**	**TTR**
it is v-ed *that*	4	**9**	44.4%	9	**24**	37.5%	8	**9**	89.0%
it V *like*	3	**19**	15.8%	4	**18**	22.2%	2	**9**	22.0%
pl-n V *together*	18	**36**	50.0%	5	**16**	31.3%	13	**27**	48.1%
there V n	9	**591**	1.5%	5	**512**	1.0%	5	**486**	1.0%
V *about* n	48	**242**	19.8%	42	**178**	23.6%	41	**174**	23.6%
V *across* n	6	**9**	66.7%	–	–	–	5	**5**	100.0%
V *after* n	4	**15**	26.7%	1	**3**	33.3%	10	**12**	83.3%
V *against* n	24	**45**	53.3%	13	**61**	21.3%	26	**61**	42.6%
V *ahead*	3	**6**	50.0%	–	–	–	5	**6**	83.3%
V *among* n	4	**5**	80.0%	5	**5**	100.0%	8	**11**	72.7%
V *around* n	10	**14**	71.4%	10	**15**	66.7%	11	**24**	45.8%
V *as if*	10	**20**	50.0%	7	**10**	70.0%	8	**14**	57.1%
V *as* n	30	**56**	53.6%	30	**100**	30.0%	64	**134**	47.8%
V *between* n	14	**22**	63.6%	11	**19**	57.9%	17	**27**	63.0%
V *down* n	33	**54**	61.1%	4	**5**	80.0%	21	**54**	38.9%
V *for* n	91	**338**	26.9%	78	**258**	30.2%	137	**386**	35.5%
V *in favor of* n	4	**7**	57.1%	3	**4**	75.0%	6	**15**	40.0%
V *in* n	165	**556**	29.7%	163	**647**	25.2%	302	**997**	30.3%
V *into* n	62	**175**	35.4%	25	**55**	45.5%	68	**181**	37.6%
V *like* n	1	**1**	100.0%	2	**2**	100.0%	12	**42**	28.6%
V n *against* n	24	**34**	70.6%	10	**20**	50.0%	20	**34**	58.8%
V n *amongst* n	7	**14**	50.0%	11	**12**	91.7%	12	**25**	48.0%
V n *around* n	6	**9**	66.7%	7	**7**	100.0%	6	**7**	85.7%
V *not.*	3	**7**	42.9%	3	**17**	17.6%	4	**19**	21.1%
V *of* n	35	**149**	23.5%	44	**100**	44.0%	59	**181**	32.6%
V *off* n	19	**32**	59.4%	4	**4**	100.0%	13	**17**	76.5%
V *over* n	29	**45**	64.4%	6	**6**	100.0%	36	**54**	66.7%
V refl pronoun	110	**207**	53.1%	87	**148**	58.8%	123	**227**	54.2%
V *so.*	9	**34**	26.5%	5	**26**	19.2%	6	**29**	20.7%
V *through* n	26	**40**	65.0%	15	**20**	75.0%	40	**72**	55.6%
V *towards* n	11	**14**	78.6%	3	**3**	100.0%	12	**15**	80.0%
V *under* n	11	**14**	78.6%	9	**18**	50.0%	20	**37**	54.1%
V *way* prep	20	**36**	55.6%	1	**1**	100.0%	5	**7**	71.4%
V *with* n	111	**307**	36.2%	97	**269**	36.1%	115	**369**	31.2%

'V *among* n', and 'V n *around* n'. Given how common the 34 selected VACs are in general English language usage, these low frequencies seem somewhat surprising. In terms of productivity of the high-frequency VACs, 'V reflexive pronoun' and 'V *with* n' are much more productive than 'V *about* n' and especially '*there* V n'. That suggests that the latter two are much more selective in the types of verbs they allow (only five different verbs in '*there* V n' in LOCNESS) than the former two VACs. In the spoken corpora we observed even lower type and token

numbers than in ICLE and LOCNESS (see Table 4). Given the smaller corpora sizes, this is perhaps unsurprising. Several VACs are not at all attested in LINDSEI_German, LINDSEI_Spanish, and LOCNEC (e.g., 'V *ahead*' and 'V n *against* n'); others are rarely used (e.g., 'V *around* n', 'V *not*.'). Only four VACs — '*there* V n', 'V *about* n', 'V *in* n', and 'V *with* n'—produced token numbers robust enough for a quantitative analysis (i.e., a relatively high number of VACs were produced across all three speaker types for each of these VACs).

Table 4: Verb type and token frequencies across VACs in spoken corpora.

	LINDSEI_German			LINDSEI_Spanish			LOCNEC		
VAC	**Types**	**Tokens**	**TTR**	**Types**	**Tokens**	**TTR**	**Types**	**Tokens**	**TTR**
it is v-ed *that*	–	–	–	–	–	–	–	–	–
it V *like*	3	**46**	6.5%	4	**97**	4.1%	3	**143**	2.1%
pl-n V *together*	5	**5**	100.0%	9	**13**	69.2%	4	**6**	66.7%
there V n	2	**165**	1.2%	3	**85**	3.5%	5	**405**	1.2%
V *about* n	22	**147**	15.0%	17	**94**	18.1%	19	**92**	20.6%
V *across* n	3	**4**	75.0%	–	–	–	1	**1**	100.0%
V *after* n	4	**5**	80.0%	2	**2**	100.0%	2	**6**	33.3%
V *against* n	–	–	–	1	**2**	50.0%	2	**3**	66.7%
V *ahead*	–	–	–	–	–	–	2	**4**	50.0%
V *among* n	–	–	–	–	–	–	1	**1**	100.0%
V *around* n	9	**18**	50.0%	4	**4**	100.0%	12	**15**	80.0%
V *as if*	3	**5**	60.0%	3	**5**	60.0%	4	**7**	57.1%
V *as* n	10	**15**	66.7%	8	**22**	36.4%	3	**8**	37.5%
V *between* n	3	**3**	100.0%	1	**1**	100.0%	4	**5**	80.0%
V *down* n	4	**4**	100.0%	2	**3**	66.7%	9	**13**	69.2%
V *for* n	33	**98**	33.7%	20	**52**	38.5%	34	**125**	27.2%
V *in favor of* n	–	–	–	–	–	–	–	–	–
V *in* n	55	**344**	16.0%	38	**256**	14.8%	39	**357**	10.9%
V *into* n	11	**33**	33.3%	3	**6**	50.0%	18	**81**	22.2%
V *like* n	8	**57**	14.0%	5	**49**	10.2%	6	**86**	7.0%
V n *against* n	–	–	–	–	–	–	1	**1**	100.0%
V n *amongst* n	–	–	–	–	–	–	1	**1**	100.0%
V n *around* n	4	**5**	80.0%	2	**2**	100.0%	4	**7**	57.1%
V *not*.	2	**5**	40.0%	3	**9**	33.3%	2	**11**	18.2%
V *of* n	4	**16**	25.0%	4	**10**	40.0%	6	**9**	66.7%
V *off* n	5	**5**	100.0%	1	**3**	33.3%	13	**30**	43.3%
V *over* n	4	**6**	66.7%	–	–	–	11	**18**	61.1%
V refl pronoun	22	**33**	66.7%	16	**28**	57.1%	31	**47**	66.0%
V *so*.	6	**28**	21.4%	6	**44**	13.6%	4	**28**	14.3%
V *through* n	5	**7**	71.4%	4	**5**	80.0%	11	**33**	33.3%
V *towards* n	–	–	–	–	–	–	3	**3**	100.0%
V *under* n	1	**2**	50.0%	–	–	–	1	**2**	50.0%
V *way* prep	1	**1**	100.0%	–	–	–	4	**6**	66.7%
V *with* n	26	**108**	24.1%	27	**117**	23.1%	27	**125**	21.6%

Overall, the inconsistency in token numbers across all learner corpus data sets underscores the need to include data from an additional source in the analysis of learner VAC knowledge. Because of the way our verbal fluency task is designed, we would expect much more homogeneous token numbers across VACs. And indeed, as Table 5 shows, that is the case. Token numbers are

Table 5: Verb type and token frequencies across VACs in verbal fluency tasks (survey).

	Survey_German			Survey_Spanish			Survey_NS		
VAC	**Types**	**Tokens**	**TTR**	**Types**	**Tokens**	**TTR**	**Types**	**Tokens**	**TTR**
it is v-ed *that*	240	**478**	50.2%	282	**614**	45.9%	328	**586**	56.0%
it V *like*	133	**678**	19.6%	196	**678**	28.9%	275	**963**	28.6%
pl-n V *together*	176	**864**	20.4%	202	**853**	23.7%	251	**1,115**	22.5%
there V n	141	**559**	25.2%	158	**416**	38.0%	159	**500**	32.0%
V *about* n	144	**454**	31.7%	142	**436**	32.6%	255	**740**	34.5%
V *across* n	161	**582**	27.7%	105	**478**	22.0%	277	**1,088**	25.5%
V *after* n	166	**529**	31.4%	146	**506**	28.9%	320	**950**	33.7%
V *against* n	159	**512**	31.1%	156	**473**	33.0%	243	**779**	31.2%
V *ahead*	117	**477**	24.5%	140	**448**	31.3%	219	**762**	28.7%
V *among* n	142	**410**	34.6%	154	**445**	34.6%	303	**866**	35.0%
V *around* n	154	**559**	27.5%	155	**536**	28.9%	317	**999**	31.7%
V *as if*	141	**567**	24.9%	196	**622**	31.5%	286	**896**	31.9%
V *as* n	215	**479**	44.9%	187	**521**	35.9%	358	**879**	40.7%
V *between* n	183	**555**	33.0%	153	**511**	29.9%	312	**937**	33.3%
V *down* n	151	**658**	23.0%	187	**603**	31.0%	256	**928**	27.6%
V *for* n	192	**520**	36.9%	156	**464**	33.6%	274	**866**	31.6%
V *in favor of* n	166	**432**	38.4%	154	**428**	36.0%	175	**518**	33.8%
V *in* n	217	**617**	35.2%	183	**584**	31.3%	318	**990**	32.1%
V *into* n	168	**615**	27.3%	135	**537**	25.1%	290	**996**	29.1%
V *like* n	151	**651**	23.2%	185	**668**	27.7%	275	**891**	30.9%
V n *against* n	145	**443**	32.7%	190	**462**	41.1%	240	**686**	35.0%
V n *amongst* n	177	**396**	44.7%	154	**445**	34.6%	252	**608**	41.4%
V n *around* n	146	**449**	32.5%	155	**536**	28.9%	252	**713**	35.3%
V *not*.	109	**479**	22.8%	98	**581**	16.9%	120	**619**	19.4%
V *of* n	119	**293**	40.6%	147	**290**	50.7%	166	**435**	38.2%
V *off* n	131	**475**	27.6%	113	**481**	23.5%	283	**912**	31.0%
V *over* n	174	**577**	30.2%	157	**594**	26.4%	318	**1,024**	31.1%
V refl pronoun	248	**573**	43.3%	245	**544**	45.0%	357	**797**	44.8%
V *so*.	140	**388**	36.1%	147	**421**	34.9%	167	**488**	34.2%
V *through* n	181	**585**	30.9%	138	**508**	27.2%	317	**1,078**	29.4%
V *towards* n	163	**545**	29.9%	122	**439**	27.8%	273	**971**	28.1%
V *under* n	174	**587**	29.6%	154	**533**	28.9%	306	**995**	30.8%
V *way* prep	171	**474**	36.1%	188	**481**	39.1%	274	**704**	38.9%
V *with* n	216	**591**	36.5%	204	**590**	34.6%	342	**997**	34.3%

robust for all VACs (ranging from 290 for 'V *of* n' in Survey_Spanish to 1,115 for 'pl-n V *together*' in Survey_NS) and higher throughout when compared to the corpus datasets. Some VACs triggered particularly high numbers of responses from learners (especially '*it* V *like*', 'pl-n V *together*', 'V *down* n', and 'V *like* n'), whereas participants evidently found it more difficult to generate multiple verbs that they associate with other VACs, including 'V *of* n', 'V n *amongst* n', and 'V *so*.' Sums of verb tokens produced across participant groups (Survey_NS: 28,276; Survey_German: 18,051; Survey_Spanish: 17,726) are expectedly much lower for both learner groups than for native speakers, with German learners producing more tokens than their Spanish peers for 20 of 34 VACs. On the whole, despite being of a less natural or authentic kind, the VAC data collected in verbal fluency tasks turned out to be richer and more robust and helped supplement that retrieved from the written and spoken corpora.

3.2 Zooming in on selected VACs

Following the overview of verb type and token lists across datasets, we will now present findings from a correlation analysis that compared learner and native speaker verb-VAC associations across VACs. We will do this for three VACs that produced sufficient token numbers across most datasets. For the three selected VACs, we will also complement the more quantitative approach with a qualitative comparison of frequency-ranked verb lists across datasets. The goal of both analyses is to identify the most common verb-VAC associations of L1 German and L1 Spanish learners (in response to RQ1) and to determine in what ways learners' verb-VAC associations are similar to or different from those of native English speakers (addressing RQ2). For reasons of space, we will begin with a detailed discussion of 'V *about* n' (covering a range of comparisons) but then only focus on one comparison each when we discuss the 'V *with* n' and 'V reflexive pronoun' results.

3.2.1 *'V* about *n'*

Let us first look at learner corpus evidence on L1 German and L1 Spanish learners' knowledge of the 'V *about* n' construction and how this compares to that of native speakers. We retrieved 242 instances of 'V *about* n' from ICLE_German, 178 from ICLE_Spanish, and 174 from LOCNESS (see Table 3

for type numbers). We then calculated Pearson correlations (*r*) between verbs used in this VAC by learner groups and native speakers and visualized the correlations in scatterplots. The possible range of absolute values for *r* is 0 to 1. The closer the absolute value is to 1, the stronger the correlation between two datasets. The correlation figures included in our results discussion express how much the sets of verbs produced by a group of learners (both in terms of types and tokens) overlap with the sets of verbs produced by a group of native speakers or a different group of learners.

Figure 1 contains two correlation plots comparing verbs used in the 'V *about* n' VAC in the ICLE subsets and LOCNESS. The x-axis shows the logarithmic frequency of the verb type in the VAC in native speaker writing; the y-axis displays the logarithmic frequency of the verb type in the VAC in learner writing (L1 German on the left, L1 Spanish on the right). Perfect overlap in verb choices between two groups (i.e., a correlation of 1), would place all verb labels neatly along the diagonal through the middle of the plot. Verbs that appear above the diagonal are markedly more frequent in the learner than the NS data; verbs that appear below the diagonal are markedly less frequent in the learner than the NS data. In our first two comparisons visualized in Figure 1 we have strong verb overlap and strong correlations ($r = 0.63$ and $r = 0.68$ respectively) but, as the verbs plotted above and below the diagonal indicate, there are also some differences in verb production by learners and native speakers. We see that THINK and CARE, for instance, occur comparatively more often in 'V *about* n' in ICLE_German than LOCNESS, that TALK and SPEAK occur more often in ICLE_Spanish than LOCNESS.

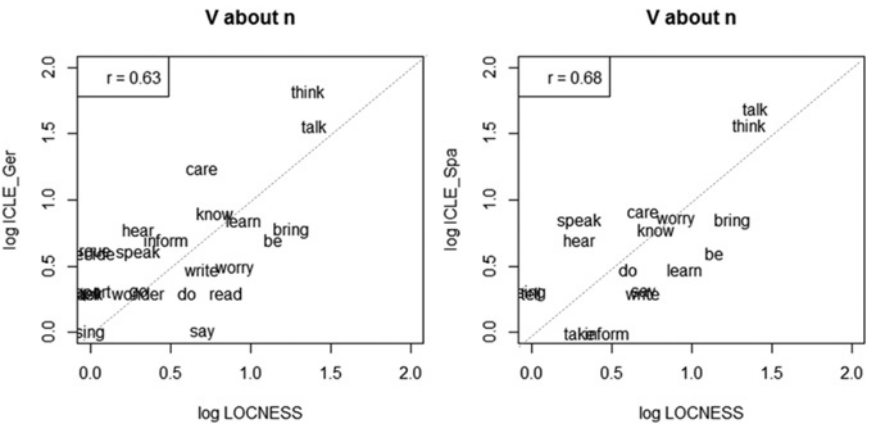

Figure 1: Correlations of verbs in learner writing (ICLE_German, left plot; ICLE_Spanish, right plot) and NS writing (LOCNESS) for 'V *about* n'.

Additionally, BRING and BE are comparatively more frequent in LOCNESS than in either of the ICLE subcorpora. A number of verbs (e.g., KNOW, LEARN, WONDER) appear close to the diagonal, which indicates that they are produced with similar frequencies by learner and native speaker writers.

Still with reference to written corpus data, we also compared verb lists across learner L1s (German vs. Spanish). Figure 2 visualizes the correlation of ICLE_German and ICLE_Spanish verbs for 'V *about* n'. Similar to the LOCNESS comparisons, we have a strong correlation ($r = 0.71$) and a number of shared verbs (plotted along or near the diagonal). Verbs preferred by German learners (plotted underneath the diagonal) include COMPLAIN and DISCUSS; verbs preferred by Spanish learners (plotted above the diagonal) include WORRY and HAVE. Verbs preferred by both learner groups include ones semantically related to communication and cognition, such as THINK and TALK. Overall, the strong *r* value and the fact that most verbs appear quite close to the diagonal suggest that, for this VAC, German and Spanish learners' verb choices are more similar than different.

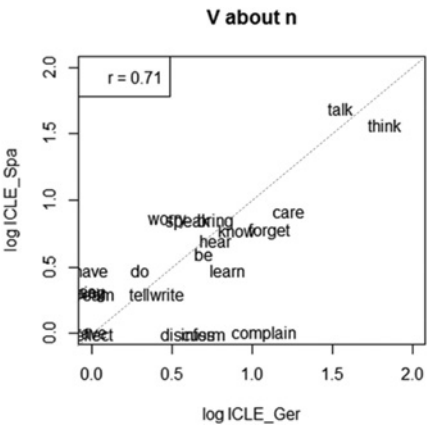

Figure 2: Correlation of verbs in German learner writing (ICLE_German) and Spanish learner writing (ICLE_Spanish) for 'V *about* n'.

Additional evidence of similarities and differences between learner and native speaker verb choices for 'V *about* n' in writing is provided in Table 6 which, for each corpus, lists the ten verb lemmas that are most frequently used in this VAC. Italicized verbs indicate overlap between learner and NS top ten lists. Of the ten verbs, seven in each learner list are shared with the NS list (including THINK, TALK, CARE, KNOW, and BE) but there are differences in token numbers

Table 6: Top ten verbs for 'V *about* n' across written corpora.

Rank	LOCNESS		ICLE_German		ICLE_Spanish	
1	TALK	25	*THINK*	67	*TALK*	49
2	THINK	23	*TALK*	36	*THINK*	37
3	BRING	18	*CARE*	17	*CARE*	8
4	BE	14	FORGET	13	*BRING*	7
5	LEARN	9	COMPLAIN	12	SPEAK	7
6	WORRY	8	*KNOW*	8	*WORRY*	7
7	READ	7	*LEARN*	7	FORGET	6
8	KNOW	6	*BRING*	6	*KNOW*	6
9	CARE	5	HEAR	6	HEAR	5
10	SAY	5	*BE*	5	*BE*	5

and rank order. The lists support what the plots in Figures 1 and 2 suggested: both learner groups strongly associate communication and cognition verbs (i.e., THINK and TALK) with this VAC. Associations with THINK appear to be particularly strong in German learners; those with TALK especially in Spanish learners. German learners produced 'COMPLAIN *about*' comparatively more often than Spanish learners and native speakers, while Spanish learners have a preference for 'SPEAK *about*'. Concordance searches of these VAC realizations in both ICLE subsets did not indicate any patterns of ungrammatical or unidiomatic uses by the learners. The verb READ, which is among the top-10 verbs in this VAC in the native speaker data, is not repeatedly used by either of the learner groups, despite similar writing tasks.

Table 7 provides the same type of verb preference data for 'V *about* n' for the three spoken corpus datasets. Type and token numbers for this VAC in the spoken corpora are lower than in the written corpora (see Table 4) but still high enough to indicate preferred speaker verb-VAC combinations. As Table 7 demonstrates, the overlap between learner and NS verb lists is similar to the written corpus lists discussed above (Table 6). Similar to the learner writing, the top two realizations of this VAC are 'TALK *about*' and 'THINK *about*', followed by 'BE *about*'. Also repeatedly used in this VAC by both learner groups is the verb COMPLAIN. We checked the LINDSEI-based 'V *about* n' concordances to see in which contexts 'COMPLAIN *about*' occurred (as we will see below, this verb is not produced in the survey data) and found that learners use it during the picture description task, in which a woman who sits for a painting does not seem happy with the result and complains about it to the painter, suggesting a possible task effect. Unique and similar to their written production, Spanish learners also show a preference for 'SPEAK *about*' in

Table 7: Top ten verbs for 'V *about* n' across spoken corpora.

Rank	LOCNEC		LINDSEI_German		LINDSEI_Spanish	
1	TALK	26	*TALK*	40	*TALK*	23
2	THINK	25	*THINK*	38	*THINK*	20
3	LEARN	7	*BE*	26	*BE*	14
4	WRITE	7	COMPLAIN	8	SPEAK	9
5	BE	5	*KNOW*	6	COMPLAIN	5
6	KNOW	4	*WORRY*	6	ARGUE	3
7	SAY	3	*LIKE*	3	*SAY*	3
8	WORRY	3	*SAY*	3	*WORRY*	3
9	LIKE	2	CARE	2	CHOOSE	2
10	BOAST	1	LAUGH	2	HEAR	2

speech. The LINDSEI_Spanish concordance for this VAC shows that in some contexts where a learner uses 'SPEAK *about*', 'TALK *about*' would actually be a better, more idiomatic fit (e.g., in "I don't wanna speak about them any more"). All other verbs attested in this VAC in LINDSEI and LOCNEC occur rather infrequently.

The verbal fluency task described above provided further insight into learner (and native speaker) verb-VAC associations. The survey datasets turned out to be richer in terms of type and token numbers than those retrieved from the learner and NS reference corpora for 'V *about* n'. This richness of data is reflected in the scatter plots displayed in Figure 3, which are considerably more populated than

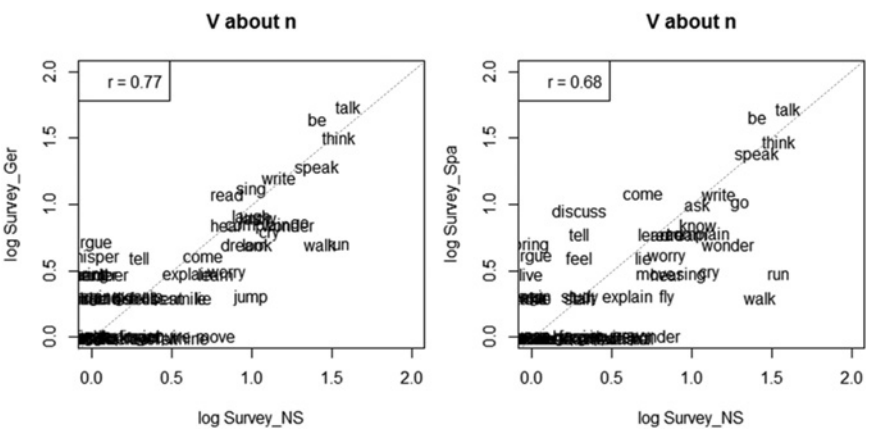

Figure 3: Correlations of verbs produced in verbal fluency tasks by learners (Survey_German, left plot; Survey_Spanish, right plot) and native speakers (Survey_NS) for 'V *about* n'.

those in Figures 1 and 2. In terms of correlations, both learner groups produced verb sets that demonstrated strong correlations with native speakers, with the L1 German group correlating with the NSs ($r = 0.77$) more strongly compared to the L1 Spanish group ($r = 0.68$). Reflecting this, a number of verbs appear along or close to the diagonal in both plots. Particularly high up towards the top right corner of the graphs we find verbs that the learner corpus analysis already highlighted as strongly associated with 'V *about* n', namely TALK, THINK, BE, and SPEAK. If we look for verbs that are produced comparatively more often by L2 participants than NSs (plotted above the diagonal), we find for example DISCUSS and TELL in the Spanish and TELL and ARGUE in the German correlation plot. All of these verbs are semantically related to one of the top 'V *about* n' verbs TALK, and thus represent creative extensions of the core semantics of the VAC which should fit the construction well but are not commonly produced by native speakers. To some L2 learners it may not be clear why 'DISCUSS *about something*' or 'TELL *about something*' are not among the preferred realizations of this VAC. These learners may not have fully figured out "what not to say" (Boyd and Goldberg 2011) and may be less sensitive to "statistical preemption" (p. 55) than native speakers, given that they have been exposed to less language input from which to derive statistical information on acceptable and unacceptable (or typical and untypical) word combinations.

If we look at the verbs plotted below the diagonal in Figure 3, indicating that they are relatively more frequent in the NS than learner data and either not at all or not commonly produced by the learners in our study, we notice verbs that belong to a different semantic set. This set includes directed motion verbs such as RUN, WALK, JUMP, and FLY. This suggests that, even at their advanced level, these learners have either not yet internalized the polysemy of this VAC or, if they have, that the 'directed motion' semantic association is much weaker for this VAC than the 'cognition/communication' one. A possible explanation for this lack of polysemy in the L1 German production data could be the fact that the German translation equivalent of the 'V *about* n' construction, '*über* n V' (e.g., *über etwas reden, über etwas nachdenken*), does not accommodate motion verbs. Such verbs would not be combined with the preposition *über* but instead with *umher* or *herum* (both translation equivalents of *around*). The top ten verb lists provided in Table 8 support the plots in Figure 3 while also providing more specific details of the verb rankings and token numbers. RUN and WALK appear high up in the NS verb production list but are absent from the learner lists. About half of the top ten verbs, including the apparently most strongly entrenched TALK, THINK, and BE, are shared across all three groups. The only 'semantic extension' verb that makes it into the top ten is DISCUSS, produced by nine Spanish learners when presented

Table 8: Top ten verbs for 'V *about* n' across survey participant groups.

Rank	Survey_NS		Survey_German		Survey_Spanish	
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

with the 's/he _____ about the... ' frame. Overall, the survey data on 'V *about* n' confirm what the corpus evidence showed on dominant verb-VAC associations while providing additional insights into the semantics of the construction.

As a final comparison for 'V *about* n', we looked at verb lists across data sources for each of the three speaker groups (L1 German, L1 Spanish, NS). Table 9 shows this type of comparison for some of the data collected from L1 Spanish learners. Italicized verbs indicate overlap across all three data types (written, spoken, experimental). While there is overlap at the top of the three lists (TALK, THINK), several verbs only appear among the top ten in a single list. We already commented on COMPLAIN in the LINDSEI-based list being the result of a task effect. Similarly, CARE and WORRY in the ICLE_Spanish list are likely

Table 9: Top ten verbs for 'V *about* n' across datasets from Spanish learners.

Rank	ICLE_Spanish		LINDSEI_Spanish		Survey_Spanish	
1	*TALK*	49	*TALK*	23	*TALK*	53
2	*THINK*	37	*THINK*	20	*BE*	46
3	CARE	8	*BE*	14	*THINK*	30
4	BRING	7	*SPEAK*	9	*SPEAK*	24
5	*SPEAK*	7	COMPLAIN	5	COME	12
6	WORRY	7	ARGUE	3	WRITE	12
7	FORGET	6	SAY	3	GO	10
8	KNOW	6	WORRY	3	ASK	10
9	HEAR	5	CHOOSE	2	DISCUSS	9
10	*BE*	4	HEAR	2	KNOW	7

related to the text type (argumentative essay) the learners were asked to produce. As the 'V *about* n' concordance from this corpus shows, learners repeatedly used phrases such as '*I do care about*', '*I don't care about*', or '*they do not worry about*' when they explained where they stood in relation to an issue they were supposed to discuss. This demonstrates that in addition to the corpus analysis providing data for only certain VACs, it also represents effects of task and text types that result in some incongruence amongst the semantic profiles of the verbs associated with a particular VAC. On the other hand, the survey verb list in Table 9 provides us with additional communication verbs and is overall semantically more coherent than the two learner corpus lists. This suggests that, in a decontextualized environment, VACs are more likely to attract the most semantically coherent verbs. In actual use and context, VACs can be influenced by task and text type and may result in data that prove difficult to interpret. However, in this case, the combination of the corpus and experimental data allowed for a fuller picture and helped confirm the preferred semantic attractions for a particular VAC.

3.2.2 *'V* with *n'*

For our second focus VAC, 'V *with* n' (a verb followed by *with*, followed by a noun or noun phrase), we will only discuss one comparison and look at learner versus native speaker writing. Figure 4 shows the correlation plots for the ICLE_German-LOCNESS and ICLE_Spanish-LOCNESS comparisons. In both cases, the learners again demonstrated strong correlations with the native speakers ($r = 0.68$ and $r = 0.7$). Based on the plot on the left, verbs used similarly frequently in this VAC in German learner and NS writing include AGREE, DO, GO, START, and INTERFERE (all plotted on the diagonal). Used comparatively more frequently by German learners than native speakers are the verbs COPE, IDENTIFY, and COMPETE; used comparatively less frequently (or not at all) are ASSOCIATE, DISAGREE, and MEET. In the ICLE_Spanish plot, AGREE is the only verb that falls on the diagonal but others such as BEGIN, START, and CONTINUE are not too far away from it, indicating similar levels of use in both datasets. Verbs that are used comparatively less often by Spanish learners than native speakers include DO, GO, and DISAGREE, while this learner group seems to use the combinations 'PLAY *with*' and 'HAPPEN *with*' more frequently than the native speakers who contributed to LOCNESS.

The top ten lists in Table 10 support many of these observations and allow us to see which additional verbs contribute to the correlation values. The number one verb across all three datasets is DEAL, making 'DEAL *with*' a

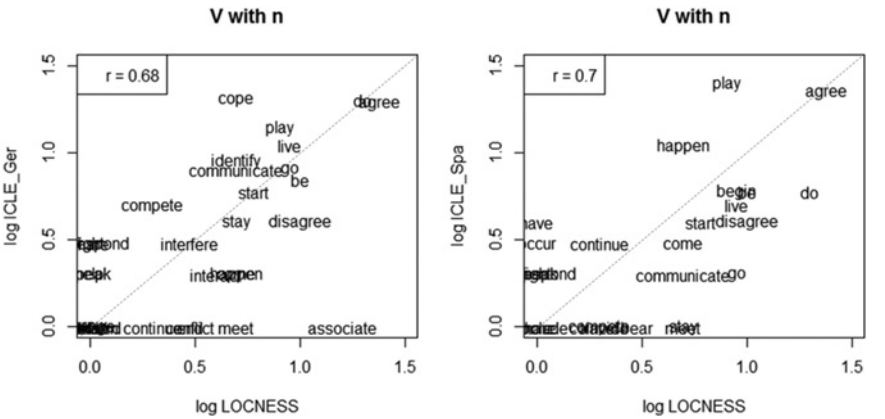

Figure 4: Correlations of verbs in learner writing (ICLE_German, left plot; ICLE_Spanish, right plot) and NS writing (LOCNESS) for 'V *with* n'.

Table 10: Top ten verbs for 'V *with* n' across written corpora.

Rank	LOCNESS		ICLE_German		ICLE_Spanish	
1	DEAL	58	*DEAL*	24	*DEAL*	31
2	AGREE	24	*DO*	20	*PLAY*	25
3	DO	20	COPE	20	*AGREE*	22
4	ASSOCIATE	16	*AGREE*	19	FINISH	12
5	BE	10	*PLAY*	14	HAPPEN	11
6	DISAGREE	10	*LIVE*	11	WORK	8
7	BEGIN	9	IDENTIFY	9	*BEGIN*	6
8	GO	9	WORK	9	MARRY	6
9	LIVE	9	COMMUNICATE	8	*DO*	6
10	PLAY	8	*GO*	8	*BE*	6

particularly strongly entrenched realization of this VAC. Other verbs that are shared between the German learner and NS lists include DO, AGREE, PLAY, and LIVE; for the Spanish learners, overlapping verbs include PLAY, AGREE, BEGIN, and DO. ASSOCIATE is a frequent NS verb that is rare or inexistent in the learner data, while WORK is comparatively more common in this VAC in learner than NS writing. Unique to the German speakers is an apparent preference for the verb COPE. A concordance analysis shows that German learners tend to use 'COPE *with*' inappropriately where verbs such as ADDRESS, DEAL (plus *with*), or TREAT would be more idiomatic choices. For instance, one learner writes '*the discovery*

of antibiotics has allowed us to cope with diseases like tuberculosis'. In this example '*treat diseases*' or '*deal with diseases*' would have adhered more closely to native speaker production. In a similar manner, a verb that only occurs in the ICLE_Spanish data for this VAC (ranked eighth in the top ten list) is MARRY, as used in '*she wants to marry with Hastings*'. This is likely the result of a cross-linguistic transfer effect (Jarvis 2013; Odlin 2013) from the learners' first language (Spanish '*casarse con*', '*to marry with*').

3.2.3 'V reflexive pronoun'

The third and final VAC we would like to share selected comparison results for is 'V reflexive pronoun', consisting of a verb followed by a reflexive pronoun such as *myself*, *yourself*, *herself*, etc. For this VAC we will only discuss the data produced by learners and native speakers in the verbal fluency tests in response to the prompts 'she _____ herself' and 'it _____ itself'. The scatterplots in Figure 5 show how similar L1 German and L1 Spanish learners' verb associations for this VAC are to those of native speakers. We observe busy plots with a variety of verbs being produced, strong correlations between the learner and NS datasets ($r = .67$ and $r = .65$ respectively), and at the same time somewhat different preferences in terms of most dominant verb associations. While associations between this VAC and verbs such as LOVE, HATE, HURT, and BE appear to be equally strong in native speakers and German learners, other verbs are produced

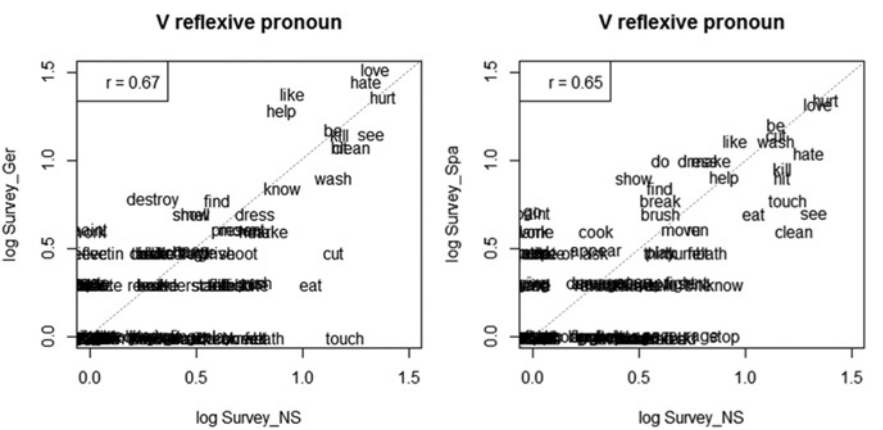

Figure 5: Correlations of verbs produced in verbal fluency tasks by learners (Survey_German, left plot; Survey_Spanish, right plot) and native speakers (Survey_NS) for 'V reflexive pronoun'.

comparatively more often in either one of the two datasets. For Survey_German these include LIKE, HELP, and DESTROY; for Survey_NS these include TOUCH, CUT, and EAT. The Spanish correlation plot also shows overlap in verb preferences (with verbs such as LOVE, HURT, BE, CUT, and HELP placed on the diagonal) as well as differences, indicated by the many verbs plotted above and below the diagonal. CLEAN, TOUCH, SEE, and EAT are in the group of verbs that native speakers produced comparatively more frequently than Spanish learners, while SHOW and DO are among those preferred by the learners. Overall for this VAC, learners' associations with physical action verbs seem to be less developed than those of native speakers.

The top ten verb lists in Table 11 provide additional information on learners' strongest verb associations for 'V reflexive pronoun', indicating a preference for emotion verbs such as LOVE, HATE, and LIKE. Among the verbs that are not shared with the NS top ten are LIKE (in both learner lists), DO and MAKE (in the Spanish list), and HELP and KILL (in the German list). German learners' strong association with HELP could be related to the fact that *helfen* (*help*) is a significant collocate of *selbst* (*self*) in German.[2] LIKE can be seen as a semantic extension of LOVE that native speakers do not think of as frequently as learners when presented with the 'V reflexive pronoun' VAC frame. Overall we again see general overlap between learners and NSs in terms of verb preferences as well as a few individual L1-specific preferences.

4 Conclusion and outlook

To summarize our findings on L1 German and L1 Spanish learners' use of selected verb-argument constructions, we observed that, as we previously found to be true for native speakers (Ellis et al. 2014a), VACs are strongly entrenched in the minds of advanced L2 English learners of two different L1 backgrounds (German and Spanish). There is considerable overlap between the verb-VAC associations of learners and native speakers, especially concerning the top-ranked, most strongly associated verbs in each construction. Differences in verb associations between NSs and learners are sometimes the results of cross-linguistic transfer from the learner's L1 (e.g., *caserse con* in Spanish), supporting our observations in Römer et al. (2014a) based on different datasets from the

2 Source: DWDS (*Das digitale Wörterbuch der deutschen Sprache*), a corpus-based dictionary of German. URL of the search: http://www.dwds.de/?qu = selbst. Selected collocate statistic: logDice.

Table 11: Top ten verbs for 'V reflexive pronoun' across survey participant groups.

Rank	Survey_NS		Survey_German		Survey_Spanish	
1	HURT	24	*LOVE*	33	*HURT*	22
2	LOVE	22	*HATE*	28	*LOVE*	21
3	SEE	21	LIKE	24	BE	16
4	HATE	20	*HURT*	23	*CUT*	14
5	CLEAN	17	HELP	19	LIKE	13
6	TOUCH	16	BE	15	*WASH*	13
7	HIT	15	KILL	14	*HATE*	11
8	KILL	15	*SEE*	14	DO	10
9	WASH	14	EXPLAIN	12	DRESS	10
10	CUT	14	*CLEAN*	12	MAKE	10

same learner populations, or related to learner creativity in the sense that L2 learners fill the verb slot in VACs with items that are semantically closely related to core verbs but less idiomatic (e.g., DISCUSS *about something*' or 'TELL *about something*'). For one of the focus VACs ('V *about* n') that attracts verbs from different semantic domains (communication, directed motion) and is used in both senses by native speakers, we also found that the learners in our study do not have the same associations and do not pick up on the different senses of this VAC. To these learners, this construction appears to be monosemous.

In addition to gaining further insights into L2 learner VAC knowledge and thus strengthening previous work by our research team on the topic and going beyond it in terms of data and VAC coverage (Ellis et al. 2014b, 2016; Römer et al. 2014a), a central purpose of our study was to assess how different data types can complement each other in researching a Construction Grammar topic in the context of second language acquisition. While being more natural and contextualized than data elicited in psycholinguistic experiments, the learner corpus data we retrieved from ICLE and LINDSEI turned out to be comparatively scarce and yielded only fairly small token numbers for the majority of selected VACs. This is in line with findings based on a smaller set of 19 VACs, reported in Römer et al. (2014b). In addition, the corpus data was susceptible to task and text type effects. Since we know that verbs in VACs display a Zipfian distribution pattern, with only a few verb types accounting for the majority of tokens and other verbs only being used once or twice (Ellis et al. 2016), larger token numbers are required if we wish to identify sets of semantically related verbs. The verbal fluency task allowed us to gather consistently high token numbers for all VACs and produced datasets robust enough for quantitative and semantic analysis. On the other hand, this type of data is arguably less natural, since participants

supply lists of verbs in a bare, decontextualized frame under time pressure, rather than while completing an actual communicative task. We could also argue that the verbal fluency task data draws more on explicit learner knowledge of VACs than especially the spoken corpus data and that it shows us what learners can consciously do when pushed to produce words that fit a certain construction.

Given the task effects we observed with ICLE and LINDSEI, however, we could also argue that the very controlled survey setting and the bare frames used as prompts were better suited to trigger more 'neutral', task-independent sets of verbs which may mirror learners' verb-VAC associations more accurately than the context-dependent VACs found in the learner corpora. In terms of the verbs that were highlighted by the two data types, we found overlap (mostly) at the top of our frequency-ordered lists (for those VACs that generated enough hits in the corpora), with the most strongly entrenched verbs being shared across datasets. However, if we had relied only on the ICLE and LINDSEI analyses, we would have missed a number of verbs that learners associate with common VACs and that are semantically related to the most frequent verbs and overrepresented verbs produced due to task and text type effects. We would be left with a more limited understanding of what learners know about VACs and the meanings they express. This is not in any way meant to diminish the value of learner corpora (ICLE and LINDSEI did provide us with valuable data on a subset of the VACs we studied) but to stress the *added* value of looking beyond corpora and considering collecting learner production data in other settings as well.

In sum then we would argue that our comparison underscores the usefulness of a combined 'corpus plus experimental data' approach. We recognize the compatibility of corpus and cognitive approaches to language analysis, believe that a bringing together of corpus and psycholinguistic evidence leads to richer results, and agree with other corpus and cognitive linguistics researchers who have acknowledged the benefits of mixed data research (e.g., Gilquin 2007, 2010; Gilquin and Gries 2009; Gries 2013; Gries and Wulff 2009; Littré 2015; Wulff 2009).

The combination of data sources and quantitative plus qualitative analytic techniques enabled us to address all of our research questions. We determined which verbs advanced German and Spanish learners of English most commonly associate with selected VACs (RQ1). We also measured how similar learners' verb-VAC associations are to those of native speakers and compared verb lists across speaker groups (RQ2). In these analyses and comparisons we drew upon corpus and psycholinguistic data and found that they complement each other well in providing insights into L2 learner VAC knowledge (RQ3).

Despite the insights we have provided, there are some limitations to our study, and more research is required to better understand what language learners know about verb-argument constructions and also how this constructional knowledge develops. Tasks on our research agenda include: adding more VACs to our list, collecting data from learners of different L1s, and tracking L2 VAC emergence by examining data produced by learners at different proficiency levels. We have begun to mine subsets of the EF-Cambridge Open Language Database (EFCAMDAT; Geertzen et al. 2013), a collection of writing samples produced by L2 learners from a range of L1 backgrounds and proficiency levels, for the VACs covered in this study. The EFCAMDAT subsets we are working with include over 40,000 texts (3.2 million words) written by Spanish learners and over 28,000 texts (2.8 million words) written by German learners at CEFR levels A1 through C2. Among the things we have analyzed so far are trends in frequency developments of verbs in VACs, dominant verb-VAC associations, type-token ratios, correlations between verbs produced by learners at different levels, and correlations between verbs produced by learners of different L1s. In addition to these written cross-sectional data, we have recently received access to a pre-release version of a new corpus of spoken learner language collected at CEFR levels B1 through C2, the Trinity Lancaster Corpus (Gablasova et al. 2017). We are currently carrying out a pilot study of a small set of five VACs in subsets of this corpus to gain additional insights on learner VAC development. We are also in the process of conducting more analyses of the verbal fluency test data, considering participants' reaction times between verbs and the orders of responses, and generating semantic networks of each VAC (including semantic relationship information; Devitt et al. 2017). So our explorations of VACs in learner production continue, and it is our hope that we have inspired others to join us in carrying out research that brings together data from different sources and, ultimately, helps us better understand how language and language acquisition work.

Acknowledgments: The authors would 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, Ben Pinkasovic, Audrey Roberson, 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

Anthony, Laurence. 2014a. *AntConc (Version 3.4.3)* [Computer Software]. Tokyo, Japan: Waseda University. Retrieved from http://www.laurenceanthony.net/.

Anthony, Laurence. 2014b. *TagAnt (Version 1.1.0)* [Computer Software]. Tokyo, Japan: Waseda University. Retrieved from http://www.laurenceanthony.net/.

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

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

Boyd, Jeremy K. & Adele Goldberg. 2011. Learning what not to say: The role of statistical preemption and categorization in a-adjective production. *Language* 87(1). 55–83.

Brown, Amanda & Marianne Gullberg. 2008. Bidirectional crosslinguistic influence in L1-L2 encoding of manner in speech and gesture. *Studies in Second Language Acquisition* 30(2). 225–251.

Brown, Amanda & Marianne Gullberg. 2010. Changes in encoding of path of motion after acquisition of a second language. *Cognitive Linguistics* 21(2). 263–286.

Cadierno, Teresa. 2008. Learning to talk about motion in a foreign language. In Peter Robinson & Nick C. Ellis (eds.), *Handbook of cognitive linguistics and second language acquisition*, 239–275). London: Routledge.

Cadierno, Teresa. 2013. Thinking for speaking in second language acquisition. In Carol A. Chapelle (ed.), *The encyclopedia of applied linguistics*. Malden, MA: Wiley-Blackwell. Accessed 13 October 2017 at http://onlinelibrary.wiley.com/doi/10.1002/9781405198431.wbeal1213/full.

De Cock, Sylvie. 2004. Preferred sequences of words in NS and NNS speech. *Belgian Journal of English Language and Literature* 2. 225–246.

Devitt, Ann, Ute Römer & Nick C. Ellis. 2017. Network analysis of knowledge of verb-argument constructions in L1 and L2 speakers. Paper presented at the American Association for Applied Linguistics (AAAL) Annual Conference, Portland, OR.

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

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

Ellis, Nick C., Matthew B. O'Donnell & Ute Römer. 2014b. Second language processing of verb-argument constructions is sensitive to form, function, frequency, contingency, and prototypicality. *Linguistic Approaches to Bilingualism* 4(4). 405–431.

Ellis, Nick C., Ute Römer & Matthew B. O'Donnell. 2016. *Usage-based Approaches to Language Acquisition and Processing: Cognitive and Corpus Investigations of Construction Grammar*. Malden, MA: Wiley.

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

Francis, Gill, Susan Hunston & Elizabeth Manning. 1996. *Grammar patterns 1: Verbs*. London: HarperCollins.

Gablasova, Dana, Vaclav Brezina, Tony McEnery & Elaine Boyd. 2017. Epistemic stance in spoken L2 English: The effect of task and speaker style. *Applied Linguistics* 38(5). 613–637.

Gass, Susan & Larry Selinker. Eds. 1983. *Language transfer in language learning*. Rowley, MA: Newbury House.

Geertzen, Jeroen, Theodora Alexopoulou & Anna Korhonen. 2013. Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT). Paper presented at the 31st Second Language Research Forum. Carnegie Mellon University, Pittsburgh, PA. 18–21 October.

Gilquin, Gaëtanelle. 2007. To err is not all: What corpus and elicitation can reveal about the use of collocations by learners. *Zeitschrift für Anglistik und Amerikanistik* 55(3). 273–291.

Gilquin, Gaëtanelle. 2010. *Corpus, cognition and causative constructions*. Amsterdam: John Benjamins.

Gilquin, Gaëtanelle, Sylvie De Cock & Sylviane Granger (eds.). 2010. *LINDSEI: Louvain international database of spoken English interlanguage*. Louvain-la-Neuve, Belgium: Presses Universitaires de Louvain.

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

Granger, Sylviane. 1996. From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In Karin Aijmer, Bengt Altenberg & Mats Johansson (eds.), *Languages in contrast. Papers from a symposium on text-based cross-linguistic studies*, 37–51. Lund: Lund University Press.

Granger, Sylviane, Estelle Dagneaux, Fanny Meunier & 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 Graeme Trousdale & Thomas Hoffmann (eds.), *The Oxford handbook of construction grammar*, 93–108. Oxford: Oxford University Press.

Gries, Stefan Th., Beate Hampe & Doris Schönefeld. 2005. Converging evidence: Bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics* 16(4). 635–676.

Gries, Stefan Th., Beate Hampe & Doris Schönefeld. 2010. Converging evidence II: More on the association of verbs and constructions. In Sally Rice & John Newman (eds.), *Empirical and experimental methods in cognitive/functional research*, 59–73. Stanford, CA: CSLI.

Gries, Stefan Th. & Stefanie Wulff. 2005. Do foreign language learners also have constructions? Evidence from priming, sorting, and corpora. *Annual Review of Cognitive Linguistics* 3(1). 182–200.

Gries, Stefan Th. & Stefanie Wulff. 2009. Psycholinguistic and corpus-linguistic evidence for L2 constructions. *Annual Review of Cognitive Linguistics* 7. 163–186.

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

Jarvis, Scott. 2011. Conceptual transfer: Crosslinguistic effects in categorization and construal. *Bilingualism: Language and Cognition* 14(Special Issue). 1–8.

Jarvis, Scott. 2013. Crosslinguistic influence and multilingualism. In Carol A. Chapelle (ed.), *The encyclopedia of applied linguistics*. Malden, MA: Wiley-Blackwell. http://onlinelibrary.wiley.com/doi/10.1002/9781405198431.wbeal0291/full (Accessed 13 October 2017).

Jarvis, Scott & Aneta Pavlenko. 2008. *Crosslinguistic influence in language and cognition*. New York: Routledge.

Kail, Robert & Marilyn A. Nippold. 1984. Unconstrained retrieval from semantic memory. *Child Development* 55(3). 944–995.

Littré, Damien. 2015. Combining experimental data and corpus data: Intermediate French-speaking learners and the English present. *Corpus Linguistics and Linguistic Theory* 11(1). 89–126.

Mollin, Sandra. 2014. *The (ir)reversibility of English binomials: Corpus, constraints, developments*. Amsterdam: John Benjamins.

Odlin, Terence. 2013. Crosslinguistic influence in second language acquisition. In Carol A. Chapelle (ed.), *The encyclopedia of applied linguistics*. Malden, MA: Wiley-Blackwell. http://onlinelibrary.wiley.com/doi/10.1002/9781405198431.wbeal0292/full (Accessed 13 October 2017).

R Development Core Team. 2012. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rebuschat, Patrick, Detmar Meurers & Tony McEnery. (eds.). 2017. Language learning research at the intersection of experimental, computational, and corpus-based approaches. [Special issue]. *Language Learning* 67(S1).

Robinson, Peter & Nick C. Ellis. (eds.). 2008. *Handbook of cognitive linguistics and second language acquisition*. London: Routledge.

Römer, Ute, Matthew B. O'Donnell & Nick C. Ellis. 2014a. Second language learner knowledge of verb-argument constructions: Effects of language transfer and typology. *The Modern Language Journal* 98(4). 952–975.

Römer, Ute, Matthew B. O'Donnell & Nick C. Ellis. 2015. Using COBUILD grammar patterns for a large-scale analysis of verb-argument constructions: Exploring corpus data and speaker knowledge. In Maggie Charles, Nicholas Groom & Suganthi John (eds.), *Corpora, Grammar, Text and Discourse: In Honour of Susan Hunston*, 43–71. Amsterdam: John Benjamins.

Römer, Ute, Audrey Roberson, Matthew B. O'Donnell & Nick C. Ellis. 2014b. Linking learner corpus and experimental data in studying second language learners' knowledge of verb-argument constructions. *ICAME Journal* 38(1). 59–79.

Slobin, Dan I. 1996. From "thought and language" to "thinking for speaking". In John J. Gumperz & Stephen C. Levinson (eds.), *Rethinking linguistic relativity*, 70–96. Cambridge: Cambridge University Press.

Wulff, Stefanie. 2008. *Rethinking idiomaticity: A usage-based approach*. London & New York: Continuum.

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