Teaming up and mixing methods: collaborative and cross-disciplinary work in corpus research on phraseology

Ute Römer[1]

Abstract

Inspired by my positive experiences gained as a member of cross-disciplinary research teams, this paper explores the value of collaborative work in corpus linguistics. I discuss selected results from three studies that showcase research on phraseology: a study that attempts to measure formulaic language in first- and second-language writing, a study on attended/unattended *this* and its patterns in student writing, and a study on speaker knowledge and use of verb–argument constructions. My collaborators on these studies include a psycholinguist, a computational linguist, a cognitive linguist and a genre analysis expert. This paper highlights the ways in which combining research methods from different fields can be beneficial to research outcomes in phraseology, and calls for more collaboration between corpus linguists and scholars from neighbouring disciplines.

Keywords: Corpus and psycholinguistic evidence, formulaic language, (un)attended *this*, verb-argument constructions.

1. Introduction

Over the past few decades, corpus linguists have made a significant contribution to moving phraseology research from the periphery to the heart of linguistics (to use Ellis's, 2008, terminology). The work of John Sinclair (e.g., Sinclair, 1991, 2004, 2008), and other researchers inspired by the British contextualist tradition, has been particularly influential in this context (e.g., Hoey, 2005; Hunston, 2002; Scott and Tribble, 2006; Stubbs, 2001; and Tognini Bonelli, 2001). By providing insights into language patterning

[1] Department of Applied Linguistics and ESL, Georgia State University, PO Box 4099, Altanta, GA 30302, USA.
Correspondence to: Ute Römer, *e-mail*: uroemer@gsu.edu

Corpora 2016 Vol. 11 (1): 113–129
DOI: 10.3366/cor.2016.0087

www.euppublishing.com/journal/cor

in different textual contexts using a variety of methods, these and other corpus linguists have demonstrated the extent to which lexis and grammar are inseparable and brought into focus a view of 'language as phraseology' (Hunston, 2002: 137). The range of recent publications on the topic are indicative of its centrality in our field (see, for example, Biber, 2009; Ebeling and Oksefjell Ebeling, 2013; Erman and Warren, 2000; Hunston and Francis, 2000; Römer, 2005, 2009, 2010; Wulff, 2009, and the contributions to Granger and Meunier, 2008; Meunier and Granger, 2008; Römer and Schulze, 2009, 2010; and to the first 2013 issue of the International Journal of Corpus Linguistics). Worth highlighting among these phraseology-focussed research endeavours are Sinclair's (1991) work on the Idiom Principle, Hunston and Francis's (2000) Pattern Grammar, Hoey's (2005) Lexical Priming, Biber and colleagues' Lexical Bundles research (Biber *et al*., 1999; and Conrad and Biber, 2004), and Gries and Stefanowitsch's Collostructional Analysis (Gries and Stefanowitsch, 2004; and Stefanowitsch and Gries, 2003). All of these have influenced the work of other scholars in the field and have inspired a multitude of projects that have promoted our understanding of the patterned nature of language.

Phraseology has also become a core interest of researchers in related fields such as natural language processing, cognitive linguistics, psycholinguistics, and language acquisition and instruction. Natural language processing uses the extraction of *n*-grams (fixed sequences of *n* words) in applications such as creating language models, predicting a foreign language user's first language, or improving machine translation (Kyle *et al*., 2013; Mariño *et al*., 2006; and Pauls and Klein, 2011). Within cognitive linguistics, the area of Construction Grammar – which views language as a system of constructions of different levels of complexity – is becoming increasingly popular (Bybee, 2010; Goldberg, 2006; Gries, 2003; Hoffmann and Trousdale, 2013; and Wulff, 2009). The popularity of formulaic language research in psycholinguistics and language acquisition is attested by the range of contributions on the topic in the 2012 Annual Review of Applied Linguistics which includes publications on phraseological items in first- and second-language acquisition, processing, instruction and use (Bannard and Lieven, 2012; Conklin and Schmitt, 2012; Ellis, 2012; and Paquot and Granger, 2012; see also, Ellis, 2003; Schmitt, 2004; and Wray, 2008).

In this paper, I would like to argue that successful collaborations between corpus linguists and scholars from related fields, as well as a skilful combination of analytic techniques – both quantitative and qualitative, can have a strong positive impact on progress and development in phraseology research. The argument is based on my positive experience of being involved in three collaborative research projects in phraseology. All three projects have benefitted from my collaboration with researchers from neighbouring disciplines, including a computational linguist, a genre expert, a psycholinguist and a cognitive linguist. They include:

(*1*) A study that attempts to measure formulaic language in corpora of academic writing by native and non-native speakers at different

proficiency levels and using a variety of operationalisations of formulaic language (O'Donnell *et al*., 2013);

(*2*) A study that combines quantitative and qualitative approaches to the distribution of attended and unattended *this* and the patterns it forms in advanced student writing across disciplines (Wulff *et al*., 2012); and,

(*3*) A study that examines verb–argument constructions in language use and in speakers' minds, drawing on corpus data and psycholinguistic evidence (Ellis *et al*., 2013; and Römer *et al*., 2015).

The following sections of this paper provide overviews of these studies, highlighting the ways in which teaming up with other researchers has led to enhanced methodologies and results. The paper closes with thoughts on future avenues for collaborative cross-disciplinary research on phraseology.

2. Three collaborative case studies in phraseology

2.1 Measuring formulaicity in L1 and L2 writing

The first case-study attempts to measure formulaicity in corpora of first- (L1) and second-language (L2) writing. It is the result of a collaboration between a psycholinguist and second-language researcher (Nick Ellis), a computational linguist (Matthew O'Donnell), and a corpus linguist (myself). A detailed account of the study can be found in O'Donnell et al. (2013).

Despite the important roles that formulaic sequences play in language acquisition, processing, fluency and idiomaticity, there is little agreement over their definition and measurement. O'Donnell *et al*. (2013) adopted an experimental design (inspired by the psycholinguist on the team) and applied four corpus-analytic measures (brought to the table by the corpus and computational linguists on the team) to samples of first- and second-language academic writing by native and non-native speakers of English with different levels of experience. One goal was to examine and compare knowledge of formulas in L1 and L2 acquisition as a function of academic writing expertise (novice *versus* expert) and language background (English as L1 *versus* English as L2). Another goal was to determine how formulaic language might be operationally defined and measured, and whether different measures would produce different results. We expected to find effects of expertise (more proficient writers show more use of formulas), language background (L1 English speakers show more use of formulas than L2 English speakers) and measure applied.

We selected the following four measures: *n*-gram frequency, *n*-gram association (Mutual Information), phrase-frames, and native norm (number of items shared with the Academic Formulas List or AFL). We selected fifteen subsets of corpora that capture academic writing and cut across the factors of nativeness and expertise. The corpora were: the International Corpus of Learner English (ICLE; capturing L2, predominantly

Corpus		Texts	Tokens	Type of writing
LOCNESS		365	269,839	L1 undergraduate
ICLE	Bulgarian	302	200,905	L2 undergraduate
	Czech	258	207,739	
	Dutch	288	271,411	
	Finnish	391	275,944	
	French	460	289,699	
	German	450	234,620	
	Italian	398	226,984	
	Polish	365	235,065	
	Russian	279	232,035	
	Spanish	260	205,003	
	Swedish	371	208,408	
MICUSP-NS		62	256,314	L1 graduate
MICUSP-NNS		72	256,318	L2 graduate
HYLAND		61	256,916	L1 expert

Table 1: Corpora used in O'Donnell *et al*. (2013). (Token counts are from WordSmith Tools 5.0.)

undergraduate writing by learners of different L1 backgrounds), the Louvain Corpus of Native English Essays (LOCNESS; undergraduate student writing by English native speakers), the Michigan Corpus of Upper-level Student Papers (MICUSP; L1 and L2 advanced-level student writing), and the Hyland Corpus of published (native or near-native) academic writing. The datasets selected from these corpora are listed in Table 1. The sets of native speaker (NS) and non-native speaker (NNS) texts selected from MICUSP are all A-graded assignments submitted by graduate-level university students from a range of disciplines. Further details of the types of texts included in MICUSP, and a full list of MICUSP disciplines, can be found in Römer and O'Donnell (2011). We regard the Hyland Corpus as an 'expert' corpus and all other corpora to be 'apprentice' corpora.

Rather than applying the analytic measures to entire (sub)corpora, as listed in Table 1, we used a novel stratified sampling approach suggested by the psycholinguist and implemented by the computational linguist on the team. This type of approach is commonly used in psychology research. Using an algorithm that divides each (sub-)corpus into sub-samples while aiming for the same number of texts and the same token count in each sample, we created eight independent samples from each of the fifteen corpus datasets. This novel approach (novel, at least, in corpus linguistics) enabled us to query each (sub-)corpus eight times instead of just once. It helped us to identify potential outliers which might skew a single average value and to determine whether differences in formula measures were significant (using ANOVAs). Table 2 shows the results of applying this sampling procedure to the 460 texts in the L1 French sub-section of ICLE. Given that none

Group	Number of files	Tokens
1	56	36,154
2	58	36,154
3	58	36,154
4	56	36,155
5	59	36,154
6	59	36,154
7	57	36,155
8	57	36,154
Total	460	289,234

Table 2: Result of applying the stratified sampling procedure to ICLE French (note that token counts are calculated using a simple whitespace tokeniser)

of the texts were split, the close similarity in text numbers and tokens across the samples is remarkable. From the 120 (15×8) samples created in this way, we (*i*) recorded the numbers of *n*-grams of different lengths with a minimum frequency of three, (*ii*) recorded the numbers of *n*-grams of different lengths with Mutual Information (MI) values above a certain length-specific threshold, (*iii*) recorded the numbers of phrase-frames[2] of different lengths with a minimum frequency of three, and (*iv*) intersected the types in each *n*-gram list with those in the Academic Formulas List (Simpson-Vlach and Ellis, 2010) and recorded the numbers of types. A more detailed discussion of the selected measures and how they are connected to the goals of our research is included in O'Donnell *et al*. (2013).

To briefly summarise the main results, we found that frequency- and MI-defined formulas are both more prevalent in advanced academic writing at expert and graduate levels than in undergraduates. Writers who contributed texts to the Hyland Corpus and to MICUSP produce more *n*-grams than ICLE and LOCNESS writers. This means that more advanced academic writers make use of higher amounts of formulaic language, including both high-frequency items, such as *on the other hand*, *due to the fact that* or *to some extent*, and items above a certain MI threshold that are often more technical or discipline-specific. Both types of formulas are relevant if one wishes to master academic writing. While there was a clear effect of expertise for both measures, neither showed significant differences between L1 and L2 writers at any level. LOCNESS and MICUSP_NS writers do not produce more *n*-grams (or more *n*-grams above a certain MI-threshold) than their

[2] Phrase-frames are sets of n-grams which are identical except for one word in the same position, for example *it is * that*, capturing the 4-grams *it is clear that*, *it is possible that*, and *it is likely that*. Phrase-frames provide insights into pattern variability and indicate to what degree language items are fixed.

non-native speaker peers (ICLE and MICUSP_NNS). For phrase-frames, there were neither significant effects of expertise nor of L1/L2 status. For AFL-defined formulas, we found a strong effect of high expertise level: Hyland Corpus contributors produce significantly more formulas that overlap with the AFL than writers in all other groups. For this last measure, there were no significant differences between the other groups, neither in terms of expertise (undergraduate *versus* graduate) nor in terms of nativeness status (English L1 *versus* English L2).

Our study has shown that different operationalisations of formulaic language produce different results. This, in turn, indicates that methodological choices in corpus linguistics may have weighty consequences that researchers need to be aware of. The absence of an L1 effect implies that formulaic sequences, including academic formulas, have to be learnt by native and non-native speakers alike. A writer's formulaic language repertoire grows with their experience of writing. Our collaborative work enabled us to gain new insights into formulaic language and helped us to explore new ways in which it can be measured.

2.2 Attended/unattended *this* patterns in student writing

The second case study looks at patterns around attended and unattended *this* in a corpus of student academic writing. It was carried out in collaboration with a cognitive linguist (Stefanie Wulff) and an academic discourse and genre expert (John Swales). A detailed account of the study can be found in Wulff *et al*. (2012). The study combines quantitative and qualitative analytic approaches to determine the distribution of the determiner *this* in advanced student writing across disciplines. It addresses the question, 'What governs an academic writer's choice between attended and unattended *this*?' Taking a methodological focus, the study also explores how results from quantitative and multifactorial analyses can guide qualitative investigations. It exemplifies how combined evidence from quantitative and qualitative methods can provide a much more comprehensive picture of a linguistic phenomenon than either method could achieve alone. The corpus that the study is based on is a pre-final version of the Michigan Corpus of Upper-level Student Papers (MICUSP; O'Donnell and Römer, 2012; and Römer and O'Donnell, 2011), consisting of 810 student writing samples from sixteen disciplines and comprising around 2.3 million words.

Although *this* is one of the most frequent words in academic writing, occupying rank 11 in a MICUSP frequency list (see Römer and Wulff, 2010), the factors that determine whether it is attended by a noun, as in Example 1, or free-standing, as in Example 2, have not received much attention in corpus research.

(*1*) This observation indicates that our method is consistent.
(*2*) This indicates that our method is consistent.

Also, with respect to this language point, style guides conflict with actual usage. While Markel (2004: 229) for example suggests that '[i]n almost all cases, demonstrative pronouns should be followed by nouns', expert writers, in fact, often use demonstrative *this* without a noun phrase: Swales (2005) found that 36 percent of all occurrences of sentence-initial *this* in the Hyland Corpus of published research articles corpus were left unattended.

In order to provide clarification on this matter and to help better understand when *this* is attended or left unattended, we carried out a systematic analysis of all sentence-initial instances of *this* in 810 texts from MICUSP (5,827 instances altogether). Of these instances of sentence-initial *this*, 57 percent (3,328) were attended by a noun or noun phrase and 43 percent (2,499) were not. We carried out (*i*) a logistic regression analysis to determine the probability of attended *versus* unattended *this* on the basis of a set of predictor variables; (*ii*) a Distinctive Collexeme Analysis (DCA) to measure the level of distinctive association between verbs and (un)attended instances of *this*; and (*iii*) a phraseological pattern analysis of the most prominent *this*+verb clusters and their distribution across MICUSP disciplines, student levels (final-year undergraduate, first-, second-, and third-year graduate) and texts.

The binary logistic regression indicated that the strongest predictor for the distribution of (un)attended *this* in MICUSP is the lemma frequency of the verb, with high lemma frequency increasing the likelihood of attended *this*. The DCA provided us with more details on this strong lexical drive and helped us to identify which items were responsible for the observed verb lemma frequency effect. Verbs that are distinctively associated with attended *this* include USE, EXAMINE, FOCUS, FIND, EXPLORE and BASE. Among the verbs that are most distinctively associated with unattended *this* are BE, MEAN, LEAD, IMPLY, SEEM and ALLOW – verbs that are used in the expression of evaluation, interpretation or discussion. The latter group of verbs form clusters with the determiner (e.g., *this is* and *this means*) that show a high degree of fixedness, meaning that they rarely allow for an intervening noun. When the following verb form is *is*, sentence-initial *this* is unattended 63.5 percent of the time; with the verb form *means*, 98.4 percent of the instances of *this* are unattended (for this verb, the only attended *this* example in MICUSP is 'This size-selectiveness means…'). The results of the DCA on verbs that typically appear in unattended contexts were corroborated by an *n*-gram extraction which highlighted a large number of fixed *this*+verb clusters in MICUSP.

A closer examination of a set of frequent *this*+verb clusters (*this is*, *this means*, *this leads*, *this implies*, *this seems* and *this allows*) included an analysis of their distribution across MICUSP disciplines and student levels, as well as of their preferred textual positions. This more qualitative view on the data pointed to a number of interesting distributional trends with respect to the clusters' use in student papers from different disciplines, and their preference to occur in (or avoid) a particular section of a paragraph or text.

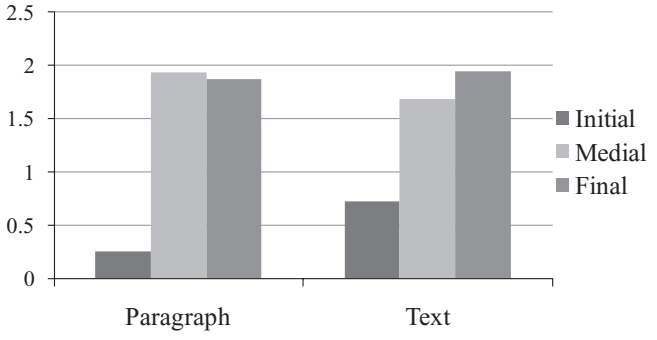

Figure 1: Distribution of sentence-initial *this seems* across paragraphs and texts in MICUSP (figures normalised per 100,000 words)

To give just one example, sentence-initial *this seems* is most frequent in Philosophy papers and, as shown in Figure 1, prefers to occur in paragraph- and text-final and -medial positions. *This seems* rarely occurs towards the beginning of a paragraph or text. Our textual distribution analysis nicely supports the DCA results on semantic groupings of verbs: combinations of *this* plus verbs which are distinctively unattended mark upcoming interpretation or evaluation. This is reflected in their positional preferences towards the end of paragraphs and texts. Correspondingly, combinations of *this* plus nouns plus verbs which are distinctively attended (according to the DCA) predominantly occur in text-initial position (Wulff *et al*., 2012: 149).

The combination of different methods and different points of view on the data – multifactorial analysis, pattern exploration and qualitative analysis of *this* in context – has enabled us to identify various hitherto unexamined properties of the attended/unattended *this* alternation. We found that it is not just the antecedent (the focus of traditional, functional studies on the topic) that may predict whether *this* is attended or not, but also the following verb. We also found that sentence-initial *this*+verb clusters form fixed contiguous sequences in MICUSP which can be regarded as meaningful units. On the whole, the results of this case study point towards an ongoing delexicalisation of *this*+verb clusters like *this is* and *this means* into textual organisation markers. These findings are inconsistent with traditional pedagogical descriptions and their claim that unattended *this* is a mere 'vague reference' which should be avoided, if at all possible. For the practice of teaching academic writing, this means that advice on (un)attended *this* found in textbooks and writing manuals appears to be over-generalised. Our study suggests that instructional materials for both native and non-native speakers of English need to recognise that certain verb forms following *this* do not favour attending nouns.

2.3 Verb–argument constructions in language use and speakers' minds

The third and final case study examines the use of English Verb–Argument Constructions (VACs; e.g., the 'V *about* n' construction illustrated by 'He thought about her suggestion') and what speakers know about the verbs that are most commonly associated with those constructions. The VAC project is an ongoing collaboration between the authors of the first case study described above: the psycholinguist, Nick Ellis; the computational linguist, Matthew O'Donnell; and myself. Detailed accounts of different parts of the project can be found in Ellis *et al.* (2013, 2014a), Römer, O'Donnell *et al*. (2014) and Römer *et al*. (2015).

The project takes a usage-based approach to VACs, their acquisition and processing. Two of its main purposes are to create an inventory of English VACs, and to determine how 'psychologically real' VACs are (i.e., how strongly they are entrenched in the speaker's mind). It addresses research questions including: how are verbs distributed across VACs? How are meanings created in VACs? What do language users know about VACs? And what role do VACs play in first and second language acquisition? All three collaborators brought specific analytical skills and methods to the project – methods from corpus linguistics, computational linguistics (especially natural language processing) and psycholinguistics / psychology. As a result, the project combines large-scale corpus analyses with different types of psycholinguistic experiments.

Starting from verb patterns identified in the COBUILD Grammar Patterns (Francis *et al*., 1996) and using tools from computational and corpus linguistics, we have mined a dependency-parsed version of the British National Corpus (BNC) for VACs. We have extracted data for around fifty VACs so far. In a team effort, and working in an iterative cycle, we have defined, searched for, reviewed and refined search patterns in order to retrieve VACs from the BNC with highest possible precision and recall (for details on the BNC-mining procedure, see Römer *et al*., 2015). We have also begun to extract VAC data from corpora of spoken and written learner English, such as the International Corpus of Learner English (ICLE) and the Louvain International Database of Spoken English Interlanguage (LINDSEI) to enable comparisons of L1 and L2 speaker use of the same VACs (Römer, Roberson, *et al*., 2014).

Psycholinguistic experiments used in this project include generative free association tasks and verbal fluency tasks. In generative free association tasks, we asked native English speakers and advanced second-language learners of different L1 backgrounds (German, Czech and Spanish) to fill forty bare VAC frames (e.g., 'she____ about the…' or 'it____ off the…') with the first verb that came to mind. In verbal fluency tasks, L1 and L2 speakers of English responded to the same forty VAC frames but were asked to produce as many verbs as they could think of in one minute. All

Figure 2: Verbal fluency task prompt for the 'V *towards* n' VAC (with responses from an advanced L1 German learner of English)

experiments were administered online using the Qualtrics survey system. The verbal fluency prompt for one of the VACs, 'V *towards* n', is displayed in Figure 2. The data collected in these experiments allows for comparisons of (*i*) BNC usage data *versus* native speaker responses, (*ii*) BNC usage data *versus* learner responses, and (*iii*) native speaker responses *versus* learner responses. The results of such comparisons allow us to determine in what ways usage influences native speaker and learner VAC production, and whether/how learner VAC knowledge differs from native speaker VAC knowledge.

Through our large-scale BNC-based corpus analyses, we demonstrated the reliability and validity of VACs in language usage. We found that VACs are Zipfian in their type–token distributions, such that a small number of verb types account for the majority of all VAC tokens while a large number of verb types only occur in a specific VAC once or twice. Figure 3 shows the distribution of verb types in the BNC for the 'V *about* n' VAC. The verbs THINK and TALK (not displayed in the figure) are the dominant verbs in this construction, each occurring more than 3,000 times, whereas verbs such as WIND and THEORIZE only have very few occurrences. The corpus analyses also indicated that VACs are selective in their verb form occupancy and coherent in their semantics (for details, see Ellis *et al.*, 2013).

The corpus findings allowed us to make predictions regarding language users' knowledge of verbs in constructions. We tested these predictions in psycholinguistic experiments. Through those experiments, we demonstrated the reliability and validity of VACs in language users' minds. We found that both native speakers and advanced second-language learners of English have strong constructional knowledge and that their VAC processing showed effects of usage frequency, contingency and prototypicality (for details, see Ellis *et al.*, 2014a, 2014b). Comparisons of experiment data collected from learners of different first languages showed that there are also systematic differences across learner groups (L1 German

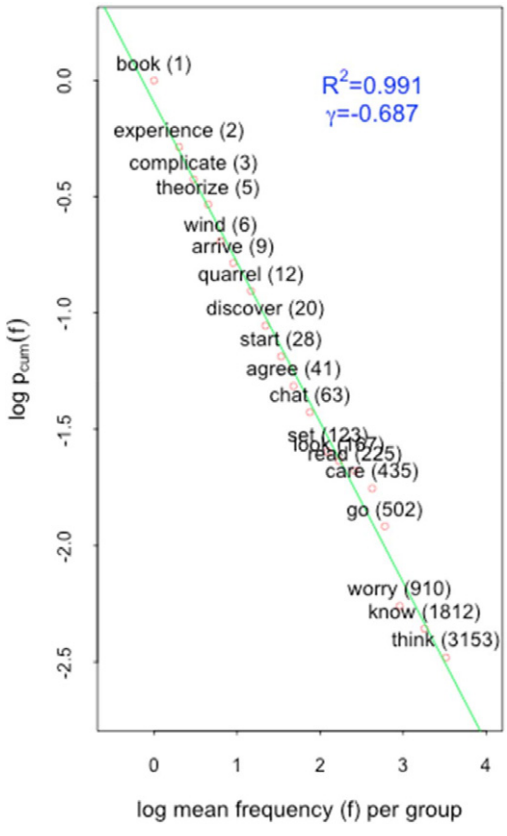

Figure 3: Verb form distribution for 'V *about* n' (based on BNC data)

versus Czech *versus* Spanish) in the associations of verbs and constructions. These differences can be explained on the basis of cross-linguistic transfer effects as well as effects of language typology that have an impact on verb semantics (Talmy, 1985). Overall, our findings suggest that learners whose first language is, like English, satellite-framed (here, German and Czech) produce more verbs that are similar to those produced by native speakers than learners whose first language is verb-framed (here, Spanish; for details, see Römer, O'Donnell *et al*., 2014). The insights into speaker knowledge and use of VACs summarised here (and discussed in the various studies referenced throughout Section 2.3) have been gained through combining methods and data sources from corpus, computational and psycho-linguistics. These insights would not have been possible with one method or data source alone.

3. Conclusion

The purpose of this paper has been to illustrate how corpus research on phraseology can be strengthened by collaborative work with scholars

from neighbouring disciplines. This was done on the basis of three case studies: a study that attempted to measure formulaic language in first- and second-language writing, a study on attended/unattended *this* and its patterns in student writing, and a study on speaker knowledge and use of verb–argument constructions (VACs). In these case studies, I teamed up with colleagues who are experts in psycholinguistics, computational linguistics, cognitive linguistics and genre analysis. All three studies have benefitted from this collaboration and from the mixing of methods, skills, techniques and insights that the individual collaborators brought to the table. In the first case study, cross-disciplinary collaboration inspired methodological enhancements in formulaic language research (including stratified sampling and formula matching) and led to new insights on the impact that different operationalisations of a concept (here, formulaicity) are likely to have on study results. In the second case study, the bringing together of quantitative and qualitative methods suggested by the collaborators resulted in converging findings and aspects of the use of *this* in academic writing that had not been reported in previous publications on the topic. In the third case study, collaboration motivated a combined-methods approach to investigating what native and non-native speakers of English know about verbs in constructions and how usage influences construction acquisition and processing. The combination of corpus and experimental methods has, in turn, produced novel results on the distribution of verbs in VACs and the factors that influence speaker VAC knowledge.

Having highlighted psycholinguists, computational linguists, cognitive linguists and genre experts as ideal candidates for team work in corpus linguistics, I would like to mention an additional type of valuable collaborator who does not usually get a great deal of recognition: the (concordance) software developer. Developers of concordance programs which are widely used by corpus researchers (e.g., AntConc, WordSmith Tools and MonoConc Pro) can have a considerable impact on our research. Unless we create our own tools, for example by writing relevant scripts in a programming language such as R (as suggested by Gries, 2009), the functionalities available in concordance packages determine what types of analyses we can or cannot carry out. To a large extent, they limit how far we can take our research. The good news is that software developers want to hear from corpus linguists and learn more about our needs. Tools are constantly evolving and are improved on the basis of user feedback. As Laurence Anthony (2009: 87), the developer of AntConc, notes, his software 'has been developed with the advice of some of the leading corpus linguists in the world, as well as the feedback and suggestions of researchers, teachers and learners.' The development continues: in 2012 and 2013, Anthony visited several corpus researchers in different parts of the world (including myself) to receive input on AntConc and on how it could be improved – for example, by adding new functionalities. In my experience, other developers of corpus analysis tools are similarly open to suggestions from users. My suggestion would be to engage more in conversation with them.

I hope that this paper has shown that collaborative, cross-disciplinary work on language phenomena can contribute to methodological enhancements in our research and that it has the potential to inspire insights that would not be possible without it. I agree with McEnery and Hardie (2012: 227) that, in corpus linguistics, 'the way ahead is methodological pluralism'. Combinations of methods and data types, such as evidence from corpora and psycholinguistic experiments, are likely to lead to richer results than one method or data source alone. Looking at a phenomenon from multiple angles allows us to capture it more fully in all its facets. For the future of research on phraseology (and other topics in linguistics), I would like to see a more extensive mixing of research methods and a triangulation of analytic techniques which, I believe, will result in further progress, productive synergies, and new, exciting discoveries.

References

Anthony, L. 2009. 'Issues in the design and development of software tools for corpus studies: the case for collaboration' in P. Baker (ed.) Contemporary Corpus Linguistics, pp. 87–104. London: Continuum.

Bannard, C. and E. Lieven. 2012. 'Formulaic language in L1 acquisition', Annual Review of Applied Linguistics 32, pp. 3–16.

Biber, D. 2009. 'A corpus-driven approach to formulaic language in English: multi-word patterns in speech and writing', International Journal of Corpus Linguistics 14 (3), pp. 275–311.

Biber, D., G. Leech, S. Johansson, S. Conrad and E. Finegan. 1999. Longman Grammar of Spoken and Written English. London: Longman.

Bybee, J.L. 2010. Language, Usage and Cognition. Cambridge: Cambridge University Press.

Conklin, K. and N. Schmitt. 2012. 'The processing of formulaic language', Annual Review of Applied Linguistics 32, pp. 45–61.

Conrad, S. and D. Biber. 2004. 'The frequency and use of lexical bundles in conversation and academic prose', Lexicographica 20, pp. 56–71.

Ebeling, J. and S. Oksefjell Ebeling. 2013. Patterns in Contrast. Amsterdam: John Benjamins.

Ellis, N.C. 2003. 'Constructions, chunking, and connectionism: the emergence of second language structure' in C. Doughty and M.H. Long (eds) Handbook of Second Language Acquisition, pp. 33–68. Oxford: Blackwell.

Ellis, N.C. 2008. 'Phraseology: the periphery and the heart of language' in F. Meunier and S. Granger (eds) Phraseology in Language Learning and Teaching, pp. 1–13. Amsterdam: John Benjamins.

Ellis, N.C. 2012. 'Formulaic language and second language acquisition: Zipf and the phrasal teddy bear', Annual Review of Applied Linguistics 32, pp. 17–44.

Ellis, N.C., M.B. O'Donnell and U. Römer. 2013. 'Usage-based language: investigating the latent structures that underpin acquisition', Language Learning 63 (Supp. 1), pp. 25–51.

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

Ellis, N.C., M.B. O'Donnell and U. Römer. 2014b. 'Second language processing of verb–argument constructions is sensitive to form, function, frequency, contingency and prototypicality', Linguistic Approaches to Bilingualism 4 (4), pp. 405–31.

Erman, B. and B. Warren. 2000. 'The idiom principle and the open choice principle', Text 20 (1), pp. 29–62.

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

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

Granger, S. and F. Meunier (eds). 2008. Phraseology: An Interdisciplinary Perspective. Amsterdam: John Benjamins.

Gries, St.Th. 2003. 'Towards a corpus-based identification of prototypical instances of constructions', Annual Review of Cognitive Linguistics 1, pp. 1–27.

Gries, St.Th. 2009. Quantitative Corpus Linguistics with R: A Practical Introduction. London: Routledge.

Gries, St.Th. and A. Stefanowitsch. 2004. 'Extending collostructional analysis: a corpus-based perspective on "alternations"', International Journal of Corpus Linguistics 9 (1), pp. 97–129.

Hoey, M.P. 2005. Lexical Priming: A New Theory of Words and Language. London: Routledge.

Hoffmann, T. and G. Trousdale (eds). 2013. The Oxford Handbook of Construction Grammar. Oxford: Oxford University Press.

Hunston, S. 2002. Corpora in Applied Linguistics. Cambridge: Cambridge University Press.

Hunston, S. and G. Francis. 2000. Pattern Grammar. A Corpus-driven Approach to the Lexical Grammar of English. Amsterdam: John Benjamins.

Kyle, K., S.A. Crossley, J. Dai and D.S. McNamara. 2013. 'Native language identification: a key n-gram category approach' in Proceedings of the

Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 242–50.

Mariño, J.B., R.E. Banchs, J.M. Crego, A. de Gispert, P. Lambert, J.A.R. Fonollosa and M.R. Costa-Jussa. 2006. 'N-gram-based machine translation', Computational Linguistics 32 (4), pp. 527–49.

McEnery, T. and A. Hardie. 2012. Corpus Linguistics. Method, Theory and Practice. Cambridge: Cambridge University Press.

Markel, M. 2004. Technical Communication. Seventh edition. Boston: Bedford/St. Martins.

Meunier, F. and S. Granger (eds). 2008. Phraseology in Language Learning and Teaching. Amsterdam: John Benjamins.

O'Donnell, M.B. and U. Römer. 2012. 'From student hard drive to web corpus (part 2): the annotation and online distribution of the Michigan Corpus of Upper-level Student Papers (MICUSP)', Corpora 7 (1), pp. 1–18.

O'Donnell, M.B., U. Römer and N.C. Ellis. 2013. 'The development of formulaic sequences in first and second language writing: investigating effects of frequency, association, and native norm', International Journal of Corpus Linguistics 18 (1), pp. 83–108.

Paquot, M. and S. Granger. 2012. 'Formulaic language in learner corpora', Annual Review of Applied Linguistics 32, pp. 130–49.

Pauls, A. and D. Klein. 2011. 'Faster and smaller n-gram language models' in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, volume 1, pp. 258–67.

Römer, U. 2005. Progressives, Patterns, Pedagogy. A Corpus-driven Approach to English Progressive Forms, Functions, Contexts and Didactics. Amsterdam: John Benjamins.

Römer, U. 2009. 'The inseparability of lexis and grammar: corpus linguistic perspectives', Annual Review of Cognitive Linguistics 7, pp. 141–63.

Römer, U. 2010. 'Establishing the phraseological profile of a text type: the construction of meaning in academic book reviews', English Text Construction 3 (1), pp. 95–119. [Reprinted in D. Biber and R. Reppen (eds). 2012. Corpus Linguistics. Volume I: Lexical Studies, pp. 307–29. London: SAGE Publications.]

Römer, U. and M.B. O'Donnell. 2011. 'From student hard drive to web corpus (part 1): the design, compilation and genre classification of the Michigan Corpus of Upper-level Student Papers (MICUSP)', Corpora 6 (2), pp. 159–77.

Römer, U. and R. Schulze (eds). 2009. Exploring the Lexis–Grammar Interface. Amsterdam: John Benjamins.

Römer, U. and R. Schulze (eds). 2010. Patterns, Meaningful Units and Specialized Discourses. Amsterdam: John Benjamins.

Römer, U. and S. Wulff. 2010. 'Applying corpus methods to writing research: explorations of MICUSP', Journal of Writing Research 2 (2), pp. 99–127.

Römer, U., M.B. O'Donnell and N.C. Ellis. 2014. 'Second language learner knowledge of verb-argument constructions: effects of language transfer and typology', The Modern Language Journal 98 (4), pp. 952–75.

Römer, U., M.B. O'Donnell and N.C. Ellis. 2015. 'Using COBUILD grammar patterns for a large-scale analysis of verb–argument constructions: exploring corpus data and speaker knowledge' in N. Groom, M. Charles and S. John (eds) Corpora, Grammar and Discourse: In Honour of Susan Hunston, pp. 43–71. Amsterdam: John Benjamins.

Römer, U., A. Roberson, M.B. O'Donnell and N.C. Ellis. 2014. 'Linking learner corpus and experimental data in studying second language learners' knowledge of verb–argument constructions', ICAME Journal 38, pp. 115–35.

Schmitt, N. 2004. Formulaic Sequences: Acquisition, Processing and Use. Amsterdam: John Benjamins.

Scott, M. and C. Tribble. 2006. Textual Patterns: Key Words and Corpus Analysis in Language Education. Amsterdam: John Benjamins.

Simpson-Vlach, R. and N.C. Ellis. 2010. 'An Academic Formulas List (AFL)', Applied Linguistics 31 (4), pp. 487–512.

Sinclair, J.M. 1991. Corpus, Concordance, Collocation. Oxford: Oxford University Press.

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

Sinclair, J.M. 2008. 'The phrase, the whole phrase, and nothing but the phrase' in S. Granger and F. Meunier (eds) Phraseology: An Interdisciplinary Perspective, pp. 407–10. Amsterdam: John Benjamins.

Stefanowitsch, A. and St.Th. Gries. 2003. 'Collostructions: investigating the interaction between words and constructions', International Journal of Corpus Linguistics 8 (2), pp. 209–43.

Stubbs, M. 2001. Words and Phrases: Corpus Studies of Lexical Semantics. Oxford: Blackwell.

Swales, J.M. 2005. 'Attended and unattended "this" in academic writing: a long and unfinished story', ESP Malaysia 11, pp. 1–15.

Talmy, L. 1985. 'Lexicalization patterns: semantic structure in lexical form' in T. Shopen (ed) Language Typology and Syntactic Description: Grammatical Categories and the Lexicon, pp. 57–149. Cambridge: Cambridge University Press.

Tognini Bonelli, E. 2001. Corpus Linguistics at Work. Amsterdam: John Benjamins.

Wray, A. 2008. Formulaic Language: Pushing the Boundaries. Oxford: Oxford University Press.

Wulff, S. 2009. Rethinking Idiomaticity: A Usage-based Approach. London: Continuum.

Wulff, S., U. Römer and J.M. Swales. 2012. 'Attended/unattended *this* in academic student writing: quantitative and qualitative perspectives', Corpus Linguistics and Linguistic Theory 8 (1), pp. 129–57.

Your short guide to the EUP Journals Blog http://euppublishingblog.com/

A forum for discussions relating to Edinburgh University Press Journals

1. The primary goal of the EUP Journals Blog

To aid discovery of authors, articles, research, multimedia and reviews published in Journals, and as a consequence contribute to increasing traffic, usage and citations of journal content.

2. Audience

Blog posts are written for an educated, popular and academic audience within EUP Journals' publishing fields.

3. Content criteria - your ideas for posts

We prioritize posts that will feature highly in search rankings, that are shareable and that will drive readers to your article on the EUP site.

4. Word count, style, and formatting

- Flexible length, however typical posts range 70-600 words.
- Related images and media files are encouraged.
- No heavy restrictions to the style or format of the post, but it should best reflect the content and topic discussed.

5. Linking policy

- Links to external blogs and websites that are related to the author, subject matter and to EUP publishing fields are encouraged, e.g.to related blog posts

6. Submit your post

Submit to ruth.allison@eup.ed.ac.uk

If you'd like to be a regular contributor, then we can set you up as an author so you can create, edit, publish, and delete your *own* posts, as well as upload files and images.

7. Republishing/repurposing

Posts may be re-used and re-purposed on other websites and blogs, but a minimum 2 week waiting period is suggested, and an acknowledgement and link to the original post on the EUP blog is requested.

8. Items to accompany post

- A short biography (ideally 25 words or less, but up to 40 words)
- A photo/headshot image of the author(s) if possible.
- Any relevant, thematic images or accompanying media (podcasts, video, graphics and photographs), provided copyright and permission to republish has been obtained.
- Files should be high resolution and a maximum of 1GB
- Permitted file types: *jpg, jpeg, png, gif, pdf, doc, ppt, odt, pptx, docx, pps, ppsx, xls, xlsx, key, mp3, m4a, wav, ogg, zip, ogv, mp4, m4v, mov, wmv, avi, mpg, 3gp, 3g2*.