Establishing the phraseological profile of a text type

The construction of meaning in academic book reviews

Ute Römer
University of Michigan

Starting from the observation that meaning does not primarily reside in individual words but in the phrase, this paper focuses on the examination of recurring phrases in language. It introduces a new analytical model that leads corpus researchers to a profile of the central phraseological items in a selected text or text collection. In this paper, the model is applied to a 3.5-million word corpus of online academic book reviews that represents part of the specialized discourse of the global community of linguists. This demonstrates how the model facilitates the study of the occurrence and distribution of the central phraseological items in linguistic book reviews, and how it helps to determine the extent of the phraseological tendency of language.

1. Introduction: Phrases and meanings

Corpus research centres around textual patterning and aims to examine how meanings are encoded in language. A growing number of publications in corpus and applied linguistics testify to the importance of issues of phraseology and provide ample evidence for the inseparability of lexis and grammar (see, e.g., Gries 2008, Hoey 2005, Hunston 2002, Hunston & Francis 2000, Partington 1998, Römer 2005a and 2009, Scott & Tribble 2006, Sinclair 1991 and 2004, Stubbs 2001, Tognini Bonelli 2001, Wulff 2009, and the contributions to Granger & Meunier 2008, Meunier & Granger 2008, Römer & Schulze 2008 and 2009). This paper starts from the assumption that the main meaning-carrying unit in language is not the word in isolation but the phrase, i.e. a sequence of two or more words which may allow for internal variation (e.g. *it would be interesting*, *it would be very interesting*). This has been most prominently stated by John Sinclair in an abstract he wrote for the

English Text Construction 3:1 (2010), **95–119**. DOI 10.1075/etc.3.1.06rom
ISSN 1874–8767 / E-ISSN 1874–8775

"Phraseology 2005" conference in Louvain-la-Neuve, Belgium, October 2005: "the normal primary carrier of meaning is the phrase and not the word; the word is the limiting case of the phrase, and has no other status in the description of meaning." (Sinclair 2008a: 409; see also Sinclair 2004: 148) Sinclair (2008a: 408) considers the phrase "central and pivotal" in language description, and comments on the superiority of an approach that focuses on phrases, rather than isolated words. For him, "[o]ne of the great strengths of a phraseological approach is the preservation of the integrity of text for much longer than alternative approaches to description, and in turn this entails the preservation of meaning." (Sinclair, 2008b: xvii). In the same vein and working in a Sinclairian British contextualist tradition, Hunston (2002: 137) talks about a theory of "[l]anguage as phraseology", and a number of corpus researchers (including myself) nowadays carry out research based on this theory, which is largely inspired by Sinclair's (1987) Idiom Principle and related ideas. According to Sinclair (2008b: xvi), "Phraseology is the ideal point of contact between a corpus and a description, because it accepts surface phenomena, and this, initially, is what a corpus provides; no pre-processing is required, no abstractions, no information such as parts of speech added."

So, if it is true that we need phrases, or "phraseological items" (a Sinclairian notion I will be using in this paper to refer to frequently occurring contiguous and non-contiguous combinations of two or more words that express a certain meaning), to locate meaning in language, a key task for any linguist who is interested in meaning construction will be to create an inventory of phraseological items in a language and to establish its phraseological profile. Since it is well known that there is a great deal of variation across registers (see, e.g., Biber 1988, Biber et al. 1999, Hofland & Johansson 1982) and that meanings are expressed in different ways in different text types, it is very hard (if not impossible) and probably not meaningful to create such an inventory for the English (or any other) language as a whole. I therefore suggest that in our phraseological explorations we focus on subsets of language, or "restricted languages" in Firth's 1956/1968 sense, and determine how meaning creation works in selected types of language that show a specialized grammar and vocabulary, "a *micro-grammar* and a *micro-glossary*" (Firth 1956/1968: 106, emphasis in original).[1] The focus in this paper will be on the restricted language of academic book reviews in the field of linguistics as captured in a corpus of that particular text type (described in Section 2).

The central aims of this paper are to demonstrate how the phraseological profile of a text or text type can be uncovered and to present the key steps involved in creating an inventory of phraseological items in a restricted language. The novelty of the approach suggested here lies in bringing existing corpus-analytic techniques together in a new way while introducing a few new techniques and concepts along the way. In discussing a new analytical model that provides insights into the

construction of meanings in text, the paper is essentially methodological in nature. By providing selected results from a large-scale empirical study of book review language that serve to illustrate the model, however, it also gives an overview of the ways in which meanings are constructed by writers who discuss the works of other members of their academic community. It hence complements existing studies on academic book reviews that have mainly focussed on the macrostructure of reviews, on disciplinary differences, or on the use of a limited set of linguistic items (e.g. hedging devices or evaluative adjectives) in this text type (see, e.g. Gea Valor 2000, Hyland 2000, Motta-Roth 1998, Römer 2005b, Suárez-Tejerina 2005).

2. The PP model: How to establish the phraseological profile of a text or text type

The question that will be addressed in the remainder of the paper is "How can we establish the phraseological profile of a text or a text type?" I have developed an analytical model, the phraseological profile model (or "PP model" for short), that enables the researcher to create an inventory of phraseological items in a text or corpus, thus providing insights into meaning creation in the discourse. The model serves to summarize the underlying procedure of text/corpus analysis and consists of four central steps: (1) the *identification* of phraseological items, (2) the determination of item-internal *variation* (e.g. A*CD and AB*D for the 4-word item ABCD, where * indicates that any item can appear in this position), (3) the examination of *functions* of the identified items, and (4) the analysis of item *distribution* across texts. The four steps will be described and discussed in the following sections (2.1 to 2.4).

The PP model is designed to be universally applicable to a wide range of spoken and written text types — including literary and non-literary texts, texts produced by native-speakers and non-native speakers of English, texts of different language varieties — and hoped to be of use to linguists, applied linguists, and literary scholars and critics alike. In the following, the model will be applied to a 3.5-million word corpus of online academic book reviews that represents part of the specialized discourse of the global community of linguists in an English-speaking context: BRILC, the Book Reviews in Linguistics Corpus. BRILC contains 1,500 reviews that were published in issues of Linguist List (see http://linguistlist.org) between 1993 and 2005. The corpus provides a good picture of how linguistic researchers worldwide discuss and assess publications in their field. For a specialized corpus of its type, BRILC is comparatively large, at least by today's standards, and serves well to represent the currently common practice in linguistic review writing in English. However, the corpus can of course not claim to be representative of

review writing in general, and certainly not of academic discourse in its entirety. It still enables us to gain insights into the language of one particular discourse community: the community of a large group of linguists around the world.

The following sections will describe the four central steps of the PP model and discuss selected findings from its application to BRILC. Section 3 will summarize a few important observations and close with some concluding thoughts on the implications of this newly developed analytical approach.

2.1 Step 1: Identification of phraseological items

In the first analytical step of the phraseological profile model, the most common phraseological items in the selected text or text collection are identified. Here a corpus-driven approach is adopted (see Römer 2005a: 7ff.; Tognini-Bonelli 2001: 84ff.) which means that, initially, items are extracted automatically from the entire corpus and the search is not limited to a pre-defined set of phrases. Corpus-driven work also implies that we work with whole texts and not text samples. Working with samples (e.g. the first 2,000 words of each text) carries the risk of missing important items that are characteristic of the text type under scrutiny but tend to occur outside the text sections covered in the samples (see Sinclair 1991: 110).[2]

In Römer (2008), I discussed whether it is possible to identify items of evaluative meaning in a corpus in any systematic way and found that, even though it may not be a straightforward task, the automatic (or semi-automatic) identification of meaningful items in a large collection of language data is actually feasible if the focus is shifted from words to phrases and phraseological search engines (i.e. corpus-analytic tools that extract groups of associated words from texts) are used. The approach described in the present paper builds on and expands the one developed in Römer (2008). The phraseological search engines used here to automatically extract recurring contiguous and non-contiguous word combinations from a corpus (here BRILC) are *Collocate* (Barlow 2004a), *ConcGram* (Greaves 2005 and 2009), and *kfNgram* (Fletcher 2002–2007).[3]

Collocate and *kfNgram* generate lists of n-grams of different lengths (i.e. sequences of n words) from a corpus, e.g. 4-grams like *as well as the* or *on the one hand* (see Figure 1 for a section of a frequency-sorted BRILC 4-gram list created with *Collocate*).[4] In addition to that, *kfNgram* also creates lists of so-called "phrase-frames" (or "p-frames"). P-frames are sets of n-grams which are identical except for one word, e.g. *at the end of*, *at the beginning of*, and *at the turn of* would all be part of the p-frame *at the * of*. P-frames hence provide insights into pattern variability and help us see to what extent Sinclair's Idiom Principle (Sinclair 1987, 1991, 1996) is at work, i.e. how fixed language units are or how much they allow

Hits	4-gram
562	on the other hand
442	at the end of
428	on the basis of
411	as well as the
356	at the same time
330	the end of the
301	of the book is
288	in the case of
268	the fact that the
226	on the one hand
216	a wide range of
215	in the context of
214	the rest of the
207	in terms of the
206	to the study of

Figure 1. Extract of a BRILC 4-gram list (*Collocate* output)

for variation.[5] Together with the types and the token numbers of the p-frames, *kfNgram* also lists how many variants are found for each of the p-frames (e.g. there are ten variants for *it would be * to*, see Figure 2). The p-frames in Figure 2 exhibit systematic and controlled variation. The first p-frame (*it would be * to*) shows that, of a large number of possible words that could theoretically fill the blank, only a small set of (mainly positively) evaluative adjectives actually do occur. In the second p-frame, modal verbs are found in the variable slot; however not all modal verbs but only a subset of them (*would*, *will*, *might*).

ConcGram allows an even more flexible approach to uncovering repeated word combinations than *Collocate* and *kfNgram* in that it automatically identifies word association patterns (so-called "concgrams") in a text (see Cheng, Greaves & Warren 2006). Concgrams cover constituency variation (AB, ACB) and positional variation (AB, BA) and hence include phraseological items that would be missed by *Collocate* or *kfNgram* searches but that are potentially interesting in terms of constituting meaningful units. Figure 3 presents an example of a BRILC-based concgram extraction, showing constituency variation (e.g. *it would be very interesting*, *it should also be interesting*). Cheng (2008: 22) defines "concgramming" as a methodology that helps identify "the 'aboutness' of a text or a corpus" which is "a product of the global patternings in the text, i.e. 'macrostructure'".

it would be * to	101	10
it would be interesting to	44	
it would be useful to	14	
it would be nice to	11	
it would be better to	9	
it would be possible to	5	
it would be helpful to	5	
it would be fair to	4	
it would be difficult to	3	
it would be necessary to	3	
it would be good to	3	
it * be interesting to	58	3
it would be interesting to	44	
it will be interesting to	8	
it might be interesting to	6	

Figure 2. Example p-frames in BRILC, together with numbers of tokens and numbers of variants (*kfNgram* output)

With all three tools (*Collocate*, *kfNgram*, *ConcGram*) I defined searches for items or patterns of different sizes. I used spans of n = 2 to n = 7 for the n-gram extractions and the p-frame generation based on the n-gram lists. *ConcGram* searches were carried out for two to four associated words. I used words that had been identified as core members of one or more of the high-frequency n-grams or p-frames (e.g. *it* + *not* + *clear*) as seed words for user-specified concgram searches. That means that I applied a serial method of analysis, moving from the corpus to n-grams, to p-frames, to concgrams — rather than going from corpus to n-grams, from corpus to p-frames, and from corpus to concgrams.[6] For each concgram, I then looked at the concgram configurations (an option in the *ConcGram* "statistics" menu) and identified if there were any frequently occurring positional or constituency variants that had not been captured by *Collocate* and *kfNgram* searches [e.g. *is not* * * *clear*]. The output files of the three tools are *candidate* lists of phraseological items that need to be manually inspected and filtered for interesting and meaningful items. As Stubbs (2007: 181) rightly notes, "there is no purely automatic way of identifying phrasal units of meaning." For example, items like *in which the*, *of the book the*, *the other hand the*, or *of the* *, which were identified by *Collocate* and *kfNgram* as frequent n-grams/p-frames, were deleted from the candidate lists because they do not constitute meaningful units (while items such

```
67  pon are backward anaphora, and it would be    interesting to see how his theory can
68  spective of grammaticalisation it would be very interesting to have a survey of the
69   semantic transparency; again, it would be very interesting to see this pursued in
70  from a theoretical standpoint, it would be very interesting to expand this analysis
71  r future research, noting that it would be especially  interesting to follow the
72  ift from OV to VO in English.  It would be particularly interesting to see if this
73  ook as exciting as I had hoped it might be, although Part 4 was quite interesting,
74  d very elegantly in the paper, it  would be interesting to discuss the
75  s in semantics. In my opinion, it  would be interesting to see how this ontological
76  oun derivatives are discussed, it  would be interesting at least to mention verbal
77  felt most positively" (p. 22). It should be noted that some interesting results
78  rs also prove a pumping lemma. It woul!  d be interesting to see further
79  pear in Linguist List reviews: it wouldn't be very interesting, I didn't make a
80   is given on this work, though it seems to be very interesting for the linguist's
81   and confined to the endnotes. It would also be interesting to set Hornstein's view
82  n of a book title was omitted. It would also be interesting to see if some of the
83  erative work on corpora. Maybe it would also be interesting to test the analyses in
84   second definition. Of course, it would also be interesting to find out that
85  ub-entries, for instance). So, it should also be interesting to find, among the
86  olved in dictionary-making and it should also be interesting to all dictionary
87  n a constituent and its copy.  It might however be interesting to seek a connection
88  ity and their self-perception. It might prove to be interesting to compare the
89  iteria seem fairly reasonable. It would, however, be interesting to study the
90  is, rhetoric, semantics, etc.  It would certainly be very interesting to see what
91  nages to carry out the action. It would most certainly be interesting to look at
92  CTIC THEORY" by Alison Henry). It seems to me that it would be  interesting to
```

Figure 3. Word association pattern (concgram) of the items *it* + *be* + *interesting* in BRILC (*ConcGram* output; sample)

as *of the book* or *on the other hand* were kept). Part of this list filtering process was a considerable amount of concordance analysis in which frequent (and potentially interesting) items from the lists were put back into their original textual contexts in order to determine whether they represented semantic units or were parts of larger units (see Figure 4 for a sample of a concordance of the 3-gram *seems to be* showing instances of the item *this seems to be*). Concordance analyses were also needed for the examination of functions described in Section 2.3 below.

In addition to basing decisions for including items in the analysis on their semantic integrity and frequency of occurrence in the selected corpus, it is desirable to use statistical tests to determine the strength of association between the components of an item (see Gries 2008). However, there is as yet no agreement on how to best measure association strengths for sequences of more than two words statistically, and I am here mainly concerned with such larger items.[7] Some pioneering work is currently being carried out in this area by corpus and computational linguists, and "lexical gravity" is put forward as a promising measure (see Gries 2009, Mukherjee & Gries 2009, O'Donnell in preparation; see also Ellis et al. 2009).

```
338    indeed, in recent years there seems to be a renewed interest in the top
339    e second case, the best thing seems to be to allow the system to learn
340    e proposition, and again this seems to be the case.  Chapter 8, Christe
341     for computer alignment! This seems to be an excellent use of computers
342    ch and writing, although this seems to be a contradiction in terms. One
343    n takes for granted (and this seems to be quite uncontroversial) that c
344    ers for this purpose and this seems to be related to a characteristic o
345     metaphysical arguments. This seems to be a significant step in the rig
346     metaphysical arguments. This seems to be a significant step in the rig
347    guistics and the Brain'. This seems to be the first serious defense of
348    onesia, Africa, etc. But this seems to be the  result of a lack of exte
349    on't speak Tok Pisin but this seems to be a real snafu.  On its own, TP
350    ntly of L2 speakers; but this seems to be the price to pay in order to
351    d pragmatic competence.  This seems to be the case even though a study
352    d pragmatic competence.  This seems to be the case even though a study
353    cality results. However, this seems to be just a parsing  failure, rath
354    antically. Nevertheless, this seems to be the case if we rely on the En
355    c, i.e. Finnish was not? This seems to be implied also on page 13. Yet,
356     act "mutually obvious". This seems to be a mistake in terminology. The
357    ase, not thematic roles. This seems to be at variance with Steinbach's
```

Figure 4. Part of a left-sorted BRILC-based concordance of *seems to be*

The described procedure of manual weeding of the *Collocate*, *kfNgram* and *ConcGram* candidate lists resulted in a database of around 8,000 phraseological items (i.e. types) that occur 20 times or more in BRILC. The results reported in this paper are based on subsets of high-frequency items (occurring at least 200 times) from this database, with a focus on items of evaluative meaning.

2.2 Step 2: Determination of item-internal variation

Following the identification of a large number of common phraseological items of different lengths in the book reviews corpus in step 1, the second step of the PP model deals with the degrees and types of internal variation that different items allow. This step examines how variable (or how fixed) a repeatedly occurring sequence of words is, where in an item variation occurs, and what the most frequent variants in a variable slot are in a non-contiguous word sequence (i.e. an n-gram with a flexible slot).

As described above, the tool *kfNgram* is able to automatically extract from BRILC lists of phrase-frames (p-frames) and their variants, i.e. sets of n-grams which are identical except for one word in the same slot (e.g. the p-frame *on the * hand* covers the n-grams *on the one hand* and *on the other hand*; *one* and *other* are listed as repeatedly occurring variants for this p-frame). In step 2 of the PP model the focus is on p-frames and their variants. Although candidate lists of p-frames

were extracted with *kfNgram* and "p-frame" is a concept used in this software (and in Fletcher's Phrases in English [PIE] database, http://pie.usna.edu), my use of the term is somewhat different from Fletcher's. While *kfNgram* treats all variations of an n-gram with a single variable slot in any position of the n-gram as p-frames (e.g. *BCD, A*CD, AB*D and ABC* for the 4-gram ABCD), I only consider n-grams with an internal variable slot (i.e. A*CD and AB*D for the 4-gram ABCD) to be p-frames. For me, sequences such as *BCD or ABC* do not constitute frames (and hence do not represent units that are of interest in the context of p-frame analysis). The status of items like *BCD and ABC* is unclear; they could either be described as 3-grams (BCD and ABC) with a preceding or following variable slot or they could be part of a larger frame, e.g. a 5-p-frame A*CDE or a 6-p-frame ABC*EF.[8] To give two concrete examples, the item *in order to* * (a p-frame in Fletcher's but not in my terminology and an item in the BRILC-based 4-p-frame list) turns out to be a component of the 5-p-frame *in order to * the* (with the * slot most frequently filled by the words *explain*, *determine*, *explore* and *understand*). Similarly, the item * *of the book*, also highlighted by *kfNgram* as a 4-p-frame with an initial variable slot, forms part of the 6-p-frame *at the * of the book*, with *end*, *beginning* and *back* filling the blank. The following analyses thus only deal with 'proper' phrase-frames in BRILC, i.e. n-grams with a variable slot in medial (not in initial or final) position. They cover p-frames of spans 3 to 6 (7-p-frames turned out to be rather rare in this corpus).

Let us first look at the different types of p-frames identified as meaningful units in BRILC and hence included in the phraseological items database. Table 1 lists all possible p-frame types by span (n = 3, n = 4, n = 5, n = 6) together with examples from BRILC. For each n, there are n-2 types of p-frames: one type for 3-p-frames, two types for 4-p-frames, three for 5-p-frames and four for 6-p-frames. All possible types have commonly occurring representatives in BRILC, e.g. we find the type AB*DE realized as *in the * of the* and *the first * of the*, among other items. Not all variations of BRILC 4-, 5- and 6-grams are equally frequent, though. Figure 5 illustrates the distribution of different p-frame types for p-frames of spans four to six. It shows that 4-p-frames of the types A*CD (49.05%) and AB*D (50.95%) are roughly equally frequent in terms of the number of different frames that are included in our database. Of the three 5-p-frame options, AB*DE is the most common type with 37.29%. For 6-p-frames, the most frequent realization is ABCD*F (31.03%), followed by AB*DEF (28.74%). Frames of the type ABC*EF are comparatively rare in our dataset (13.79%).

Findings like these do not only help us better understand the phraseological profile of a text type, they are also potentially interesting for distinguishing one text type (or register) from another. Biber (2009a: 294), for example, refers to differences between English conversation and academic writing with respect to the

ways in which a sequence of words is fixed and where in the sequence variable slots are most likely to occur (see also Biber 2009a). He, however, only looks at a selection of 4-word "lexical bundles" (n-grams) and their internal variation, including variation of the *BCD and ABC* kind. It hence may be the case that some of the register differences Biber highlights (e.g. the higher frequency of ABC* sequences in conversation as compared to academic writing) are in fact due to varying numbers of occurrence of items of different spans (e.g. perhaps higher frequencies of 5-grams and thus of ABC*E sequences in conversation).

Table 1. Types of p-frames in BRILC, sorted by span

Span	P-frame types	Examples from BRILC
3	A*C	*the * of, a * job*
4	A*CD	*the * that the, the * of language*
	AB*D	*at the * of, on the * hand*
5	A*CDE	*the * of the book, at * end of the*
	AB*DE	*in the * of the, the first * of the*
	ABC*E	*the book is * into, it would be * to*
6	A*CDEF	*the * part of the book, the * of the book is*
	AB*DEF	*at the * of the book, the first * of the book*
	ABC*EF	*in the first * of the, the book is * into three*
	ABCD*F	*it would have been * to, book is divided into * parts*

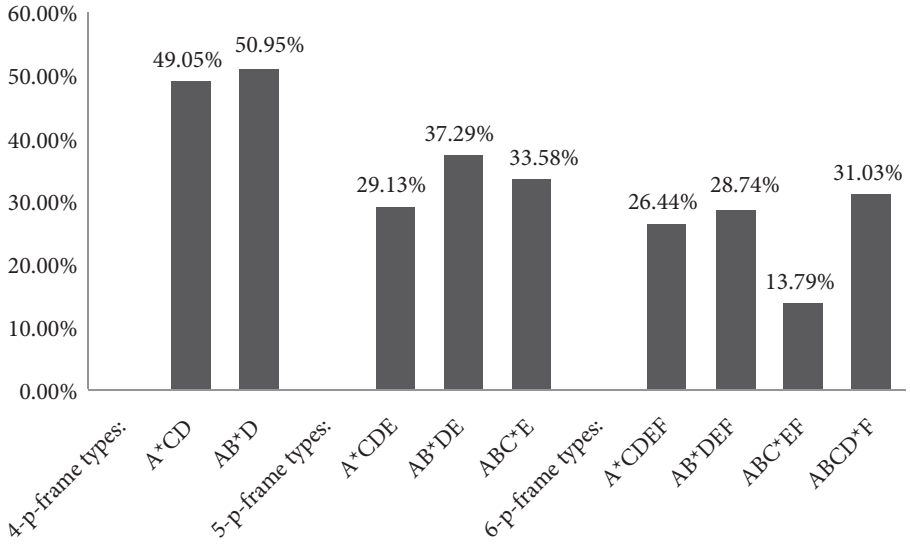

Figure 5. The distribution of different p-frame types for p-frames of spans 4, 5 and 6

In order to find out more about the fixedness of phraseological items, step 2 of the PP model not only examines the *types* but also the *degrees* of internal variation of common word sequences in a text or text collection. The model introduces a new measure to help capture how much variation an n-gram allows: the variant/p-frame ratio (VPR). The VPR captures the relation of different words that fill the blank (*) slot in a p-frame to the number of p-frame tokens. It functions like a type/token ratio for frames. If a p-frame occurs 1,000 times in a text collection and has only two variants, it has a VPR of 0.2%; a p-frame that occurs 1,000 times but has 500 variants has a VPR of 50%. In other words, a low VPR indicates that there are only very few different variants per p-frame type and that we are dealing with a rather fixed item that does not show much variation. I determined the variant/p-frame ratios for all p-frames in the BRILC-based database that occur at least 200 times in the corpus (280 items altogether) and found an average VPR of 8.84% (median: 8.71%, standard deviation: 4.65%), with values ranging from 0.25% (found for *on the * hand*; two variants, 800 p-frame tokens) to 19.78% (found for *the * and the*; 106 variants, 536 tokens). Table 2 provides examples of p-frames with above and below average variant/p-frame ratios. We see that items like *at * end of* and *a *range of* are relatively fixed and only accept a small selection of words as blank-fillers, whereas items like *by the * of* and *discusses the * of* exhibit a high degree of variation and allow for a range of different words to fill the * slot.

Table 2. Selected BRILC p-frames with high and low variant/p-frame ratios (VPRs)

	P-frame	Token number	Variant number	VPR
High VPRs	*the * and the*	536	106	19.78%
	*by the * of*	329	58	17.63%
	*discusses the * of*	236	41	17.37%
	*about the * of*	268	44	16.42%
Low VPRs	*on the * hand*	800	2	0.25%
	*at * end of*	448	2	0.45%
	*with * to the*	341	5	1.47%
	*a * range of*	314	8	2.55%

A final question I would like to address in the context of determining phraseological item-internal variation is whether or not the distribution of variants per p-frame is Zipfian, that is, whether it follows Zipf's law (Zipf 1935) which says that the frequency of a word is inversely proportional to its rank in a corpus-based frequency table. It is characterized by a power law relation. In other words, a small number of high-frequency items (types) account for the majority of instances (tokens), and token numbers drop drastically among the first few high-frequency items. In our case this would imply that a small number of variants account for a large share of tokens of a selected p-frame. A Zipfian distribution would also entail

that the first few realizations of a p-frame in a p-frame display (see Figure 2) appear in our n-gram lists because they are highly frequent.

If the distribution of variants across p-frame tokens was Zipfian, this would have implications for language teaching and in particular the teaching of EAP because it would mean that most instances of a p-frame were covered by just teaching a handful of variants or realizations of the frame. To refer back to the first example in Figure 2, knowing only the first four variants that most commonly fill the * slot in the frame (*interesting*, *useful*, *nice*, *better*), would enable learners to capture 78 out of 101 instances of *it would be * to* in the book reviews corpus. This would be different if the variant distribution was non-Zipfian and there was a long list of variants each of which occurred only once or twice in the p-frame. I analysed a selection of p-frames from the database with high token numbers and found that (in the analysed sample) p-frames always show Zipfian distributions if they have average or above-average VPRs. For p-frames with low variant numbers (and thus low VPRs) it was obviously harder to tell whether the few variants followed a power law distribution. The graph for the p-frame *on the * of* and its variant types is displayed in Figure 6. The distribution of variants for this p-frame is clearly Zipfian, with the most frequent variant (*basis*) occurring 431 times, followed by the second most frequent variant (*part*) with 92 and the third most frequent one (*role*) with 83 occurrences (see full list of variants in Table 3). The majority of variant types were only found to occur between one and ten times. These findings suggest that the * slot may not always be as open as representations like AB*D suggest and that Sinclair's Idiom Principle is at work more often than not. Further related discussions of Zipfian distributions in language and their implications for language

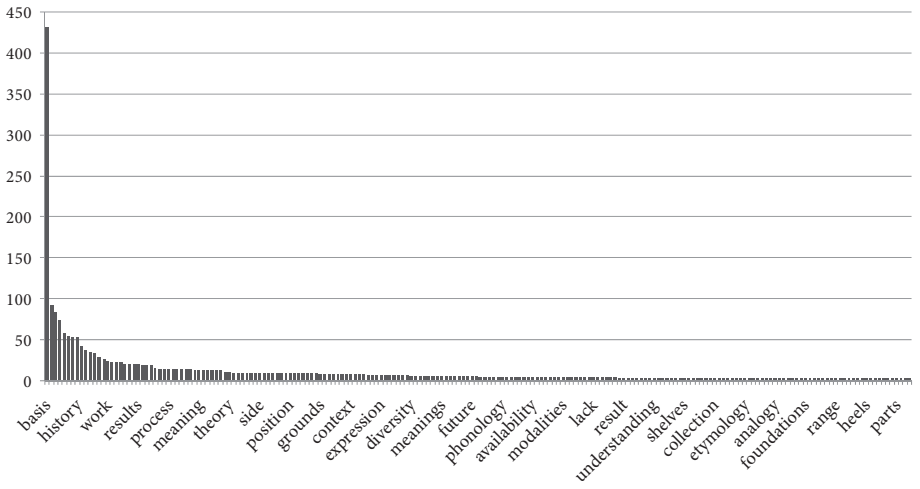

Figure 6. Variant type distribution for the p-frame *on the * of*, following Zipf's law

acquisition and language teaching can be found in Ellis 2009, Ellis & Cadierno 2009, and Ellis & Ferreira-Junior 2009.

Table 3. List of all variants that occur at least 3 times in the p-frame *on the * of*, in decreasing order of frequency

Variant	Tokens	Variant	Tokens	Variant	Tokens	Variant	Tokens
basis	431	*perception*	9	*politics*	4	*rest*	3
part	92	*number*	9	*set*	4	*developments*	3
role	83	*typology*	9	*characteristics*	4	*domain*	3
nature	73	*field*	8	*co-occurrence*	4	*efficacy*	3
use	57	*occurrence*	8	*performance*	4	*collection*	3
development	54	*degree*	8	*phonology*	4	*consequences*	3
acquisition	52	*position*	8	*philosophy*	4	*contribution*	3
history	52	*types*	8	*roles*	4	*experience*	3
notion	41	*representation*	8	*problems*	4	*form*	3
analysis	37	*effects*	8	*purpose*	4	*formulation*	3
question	34	*theme*	8	*accuracy*	4	*elaboration*	3
concept	33	*discussion*	8	*right*	4	*etymology*	3
level	28	*impact*	8	*availability*	4	*evaluation*	3
issue	26	*grounds*	7	*background*	4	*anti-locality*	3
work	23	*state*	7	*presentation*	4	*applicability*	3
status	22	*identification*	7	*assumption*	4	*appropriate-ness*	3
importance	22	*choice*	7	*principle*	4	*achievement*	3
study	22	*construction*	7	*speed*	4	*advantages*	3
problem	20	*relation*	7	*stability*	4	*analogy*	3
structure	20	*case*	7	*modalities*	4	*behaviour*	3
topic	20	*context*	7	*dialects*	4	*bookshelf*	3
results	19	*island*	7	*model*	4	*bookshelves*	3
syntax	18	*formation*	7	*system*	4	*architecture*	3
subject	18	*discourse*	7	*surface*	4	*articulation*	3
evolution	18	*place*	6	*features*	4	*behavior*	3
interpretation	15	*kinds*	6	*languages*	4	*foundations*	3
presence	14	*effectiveness*	6	*lack*	4	*pragmatics*	3
existence	14	*expression*	6	*definition*	4	*primacy*	3
process	14	*findings*	6	*organization*	4	*principles*	3

Table 3. *(continued)*

Variant	Tokens	Variant	Tokens	Variant	Tokens	Variant	Tokens
teaching	14	*effect*	6	*idea*	4	*pattern*	3
distribution	13	*variety*	6	*notions*	4	*perspective*	3
description	13	*complexity*	6	*comparison*	4	*placement*	3
type	13	*content*	6	*territory*	3	*range*	3
origins	13	*grammatical-ization*	6	*result*	3	*relationship*	3
issues	12	*diversity*	5	*varieties*	3	*resolution*	3
meaning	12	*concepts*	5	*rise*	3	*production*	3
language	12	*selection*	5	*works*	3	*pronunciation*	3
semantics	12	*limitations*	5	*vocabulary*	3	*questions*	3
emergence	12	*spread*	5	*situation*	3	*goal*	3
interaction	12	*distinction*	5	*value*	3	*heels*	3
origin	12	*limits*	5	*understanding*	3	*identity*	3
possibility	10	*meanings*	5	*treatment*	3	*frequency*	3
theory	10	*grammar*	5	*training*	3	*function*	3
properties	9	*face*	5	*universality*	3	*genesis*	3
example	9	*implications*	5	*usage*	3	*mechanisms*	3
occasion	9	*speech*	5	*uses*	3	*necessity*	3
application	9	*relevance*	5	*topics*	3	*parts*	3
kind	9	*strength*	5	*shelves*	3	*length*	3
influence	9	*future*	5	*scope*	3	*logic*	3
side	9	*functions*	5	*standardiza-tion*	3	*mapping*	3

2.3 Step 3: Examination of functions of the identified items

We now know what the most common phraseological items in our text collection are and what types of internal variation they allow. Step 3 of the PP model will focus on the functions of the identified items. Determining what meanings are expressed by the most frequent phraseological items will help us understand *why* they are used so often by members of the discourse community in question (here a large group of linguists from around the world). In order to identify what our high-frequency items do and how they function in academic book reviews, most of them need to be looked at in the context of concordances. While it is not very hard to determine that the frequent 4-gram *it is not clear* functions as a marker of

negative evaluation, there are a large number of items that are difficult to classify in isolation. An item like *at the same time*, for example, is much harder to interpret out of context. A look at a concordance, however, shows that in book reviews the 4-gram (when used in its non-temporal sense) mostly functions to prepare the ground for positive evaluation, as in *At the same time, intriguing individual observations are raised* […]. It turned out that our phraseological items were not always mono-functional, but one meaning always dominated and outnumbered potential second or third meanings. I therefore decided to assign only one function to each item which was the dominant one expressed by the item in the selected text type. When it came to p-frames (word sequences with a variable slot), the functional labelling was obviously particularly challenging. They were hence only assigned a function when their variants (i.e. the words that fill the variable slot) were semantically related (like those in Figure 2) and the same function was valid for all realizations of the p-frame.

Given that almost every item needs to be looked at in context, step 3 is the most time-consuming and currently still ongoing part of the analysis (only about 10% of all items have so far been functionally classified). At this point, it is too early to quantify the findings on functions of phraseological items in BRILC. I will, however, make a few qualitative observations, list the functions expressed by the most frequent items in the database, and provide a few examples.

The examined BRILC n-grams and p-frames expressed altogether four different functions. Unsurprisingly (given that we are here dealing with a highly evaluative text type), a large proportion of the classified items, e.g. *it is * that*, *would be * to*, *a wide range of*, express EVALUATION, both of a positive and negative kind. Another group of items, including *in the * chapter*, *the * of the book*, *in the first part*, function to refer to the STRUCTURE of the book under review. Reference to a book's CONTENT, as in *the * of English*, *the history of*, *the relationship between * and*, was identified as a third function, and a fourth set of repeatedly occurring items (e.g. *in order to*, *with respect to*) are used to organize the DISCOURSE. These four functions (expressing evaluation, referring to a book's structure, referring to the content of a book, and organizing the discourse,) are what characterize the text type under analysis. Their examination is therefore an essential part of the creation of a text's or text type's phraseological profile.

2.4 Step 4: Analysis of the distribution of items across texts

The fourth and final step of the PP model analyzes the distribution of common phraseological items across texts in a text collection. In doing so, this step relates phraseological items with text structure and hence highlights instances of what Hoey (2005) calls "textual colligation". According to Hoey (2005:13), "[e]very

word is primed to occur in, or avoid, certain positions within the discourse; these are its textual colligations." Textual colligation has been extensively studied in the language of newspapers (e.g. Hoey 2005, 2009; Hoey & O'Donnell 2008; Mahlberg & O'Donnell 2008) but the concept is only now being applied to the analysis of academic discourse (see also Römer & O'Donnell 2009). Knowing where in a text an item most commonly occurs and which positions it avoids, facilitates text processing and is particularly important in the production of a text of the selected type. In the case of academic book reviews, for instance, it helps the reader, and especially the writer, to know which phraseological items are most commonly used in the introduction or critical evaluation section towards the end of the review.

Software tools for corpus analysis usually offer ways to observe the distribution of items across texts. Users find a "dispersion plot" function in *WordSmith Tools* (Scott 2008), a "concordance plot" function in *AntConc* (Anthony 2007), and a "distribution of hits" function in *MonoConc Pro* (Barlow 2004b). These functions provide a graphic illustration of the individual occurrences of a word or phrase across a text, usually in the form of a bar code with each occurrence represented by a thin line (see Figure 7). Such displays are very useful but evaluating them in a systematic way — other than just eyeballing long lists of bar codes — is far from straightforward. With the help of my colleague Matthew Brook O'Donnell, I therefore developed an alternative approach to systematically tracking item occurrence across the 1,500 BRILC text files. Each file was divided into four parts of equal size (in terms of number of words) and division tags were inserted around each text quarter to mark the beginning ( ) and end ( ) of a quarter. A manual analysis of a set of twenty files randomly pulled from the corpus indicated that the quarters are related to structural elements of the review in that quarter 1 tends to correspond with the introduction section of the review, quarters 2 and 3 (i.e. the middle 50%) with the summary of the book's contents, and quarter 4 with the critical evaluation/conclusion section. With this 'quartered' version of the corpus it was then possible to retrieve frequency lists of n-grams and p-frames separately for collections of the first, second, third and fourth 25% of each review. This version also allowed me to compute the shares of occurrences of the identified 8,000 phraseological items across BRILC quarters and determine, for example, what percentage of the 562 instances of *on the other hand* occur in each of the four text segments.

Table 4 shows ranked lists of the ten most frequent 4-grams in each of the four BRILC quarters (extracted with *kfNgram*). We see that five out of ten items (set in small caps in Table 4) are shared by all quarter lists: *on the other hand*, *on the basis of*, *as well as the*, *at the end of*, and *at the same time* — all items that appeared in the BRILC (all) frequency 4-gram list. These five items have made it on all top-10 lists but they do occupy different ranks in each list. *At the same time*, for example,

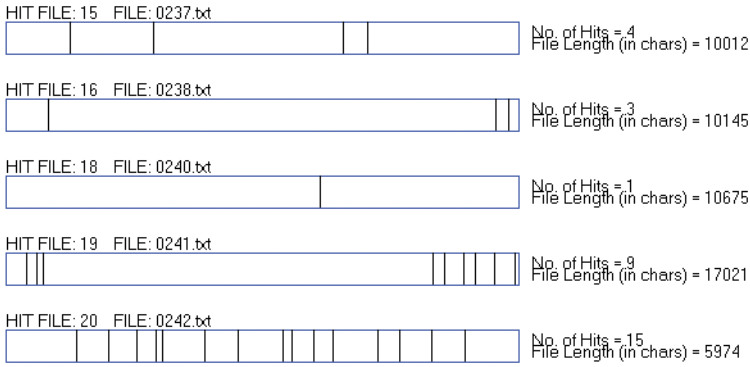

Figure 7. Part of the *AntConc* concordance plot for the 2-gram *the book* in BRILC (5 out of 1,500 files displayed)

Table 4. Top-10 4-grams across BRILC quarters

	Quarter 1	Quarter 2	Quarter 3	Quarter 4
1	*the book is divided*	ON THE BASIS OF	ON THE OTHER HAND	ON THE OTHER HAND
2	*book is divided into*	ON THE OTHER HAND	ON THE BASIS OF	AT THE END OF
3	*is a collection of*	AS WELL AS THE	AS WELL AS THE	AT THE SAME TIME
4	*an overview of the*	*the end of the*	AT THE END OF	AS WELL AS THE
5	AT THE END OF	*in the case of*	AT THE SAME TIME	*of the book is*
6	*of the book is*	AT THE END OF	*in the case of*	*the fact that the*
7	ON THE OTHER HAND	AT THE SAME TIME	*the fact that the*	*the end of the*
8	AS WELL AS THE	*the rest of the*	*the end of the*	ON THE BASIS OF
9	ON THE BASIS OF	*in terms of the*	*with respect to the*	*it would have been*
10	AT THE SAME TIME	*in the context of*	*on the other hand*	*in the case of*

appears at rank ten in the quarter 1 list, at rank seven in quarter 2, at rank five in quarter 3, and at rank three in quarter 4. This could mean that, although the item is generally common in book reviews, it seems to have a preference to appear in final text quarters. We also find a small number of unique 4-grams in each quarter list (underlined in Table 4). These include items that refer to the structure of the book under review in quarter one (e.g. *the book is divided*), discourse organizing items in quarters two and three (e.g. *in terms of the*, *with respect to the*), and an item that functions to introduce negative evaluation in quarter four (*it would have been*). These initial findings on the distribution of 4-grams across BRILC texts already offer some interesting pointers as to what types of meanings are created in which section of a book review. Let us now look at a selection of high-frequency phraseological items from the database in a bit more detail to see how they are distributed across texts and gain further insights into meaning creation in book reviews.

The items I selected for the more detailed text position analysis are *at the same time* (a 4-gram found in all top-10 quarter lists), *on the * hand* (summarizing *on the one hand* and *on the other hand*), *a * range of* (* slot filled by *wide*, *broad* and semantically related adjectives), *it would have been*, and *it is not clear.* They all occur frequently in BRILC and are included in our phraseological items database. Figures 8 to 12 illustrate the percentages of item token numbers for each text quarter and show which of the selected items favor or avoid certain textual positions.

Starting with *at the same time*, we observe a mild preference of the 4-gram for the fourth and final text quarter and an otherwise more or less even distribution (see Figure 8). The chi-square test shows that there is a 6.4% chance of this distribution occuring at random which means that the result is not actually statistically significant (unless we use a comparatively high p-value threshold). So, it appears safe to use *at the same time* in all parts of an academic book review. The 4-p-frame

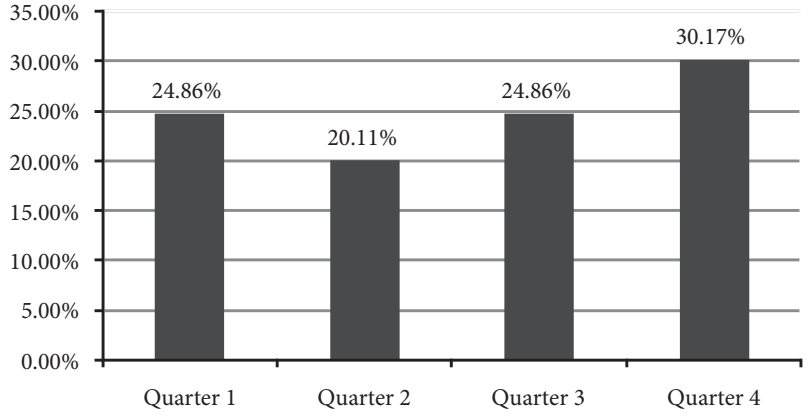

Figure 8. Distribution of the phraseological item *at the same time* across BRILC quarters

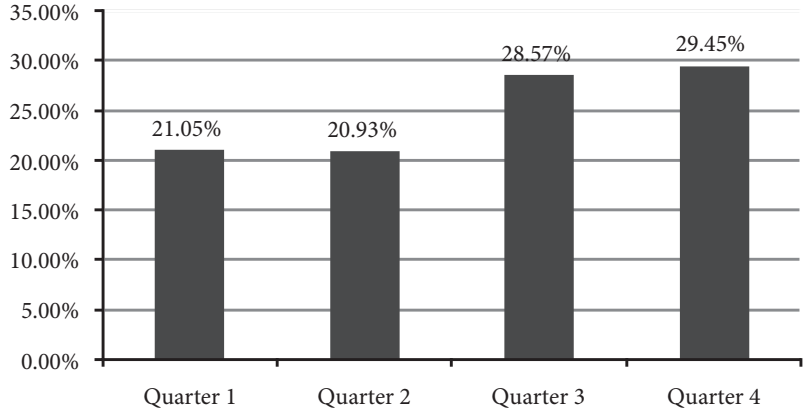

Figure 9. Distribution of the phraseological item *on the * hand* across BRILC quarters

*on the * hand* shows, as indicated in Figure 9, a preference to occur in the third and fourth quarters of BRILC reviews and a dispreference for quarters 1 and 2 (chi-square value significant at .0001 level). This could mean that writers of book reviews do not usually introduce an argument in the first 50% of their text but are more likely to do so in the second half. The distributional profile for *a * range of* is also quite interesting (and statistically significant), as Figure 10 demonstrates. This item is commonly used to express positive evaluation, as in *the book offers a wide range of useful and interesting insights into the research area of spatial language*. It is thus not surprising that *a * range of* would occur most frequently in the final quarter of BRILC reviews which generally corresponds structurally to the "critical evaluation" section of the text. However, it also occurs very often in quarter 1 (with 29.97% of all tokens) which implies that positively evaluative meanings are created in introduction sections too. Lastly, *it would have been* and *it is not clear* show very

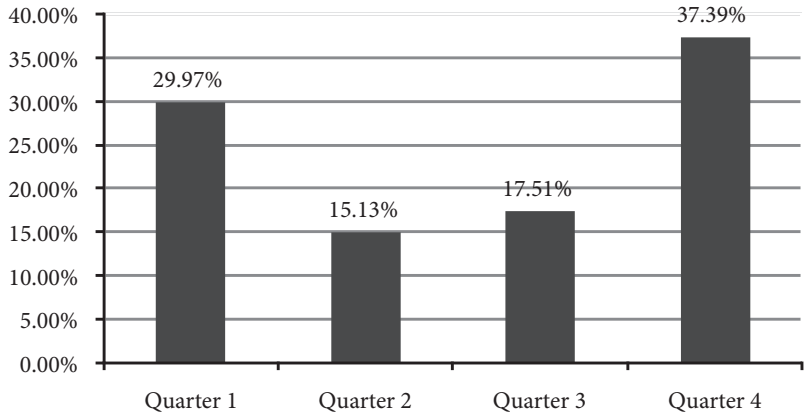

Figure 10. Distribution of the phraseological item *a * range of* across BRILC quarters

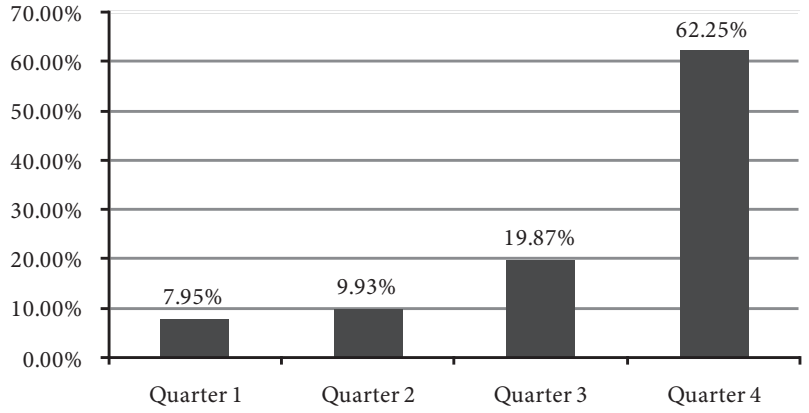

Figure 11. Distribution of the phraseological item *it would have been* across BRILC quarters

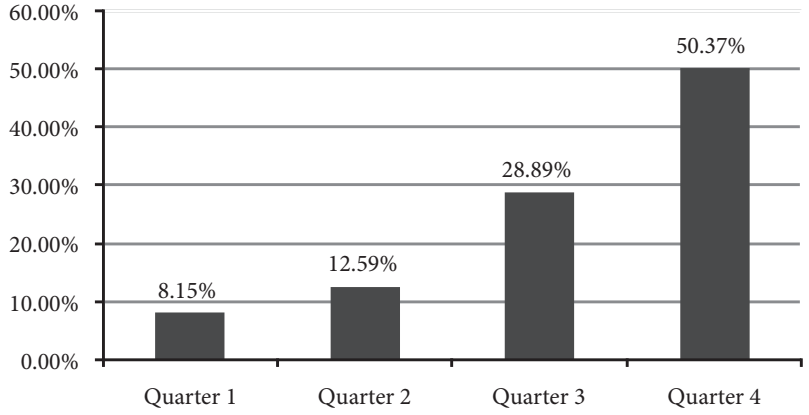

Figure 12. Distribution of the phraseological item *it is not clear* across BRILC quarters

similar distributional profiles, displayed in Figures 11 and 12. Both 4-grams are used to convey negative evaluation (as in *it would have been interesting to also compare with results in other corpus-based studies of translations*) and strongly favour final text quarters while avoiding quarters one and two (chi-square very highly significant). So, although we know that evaluation is frequently expressed not only in the final section of LinguistList book reviews but also in review introductions, items such as *it would have been* and *it is not clear* that express negative evaluation tend not to occur at the beginning of texts but cluster towards the end.

The results of our step 4 analyses have highlighted multiple instances of textual colligation (i.e. a clear preference or dispreference of phraseological items for certain positions in a text), some of which have been illustrated here. The textual colligations exemplified in the previous paragraphs have not only indicated what items tend to occur where in a text but also helped us gain a better understanding of where in a text certain kinds of meanings are mainly expressed, e.g. structure-related items in the first quarter and negatively evaluative items in the fourth quarter of the text. This kind of analysis is a central part of getting at the phraseological profile of a text and better understanding meaning creation in a particular type of discourse.

3. Summary and conclusion

This article has put forward a new model for text and corpus analysis: the PP model. The paper has demonstrated how the model facilitates the study of the occurrence and distribution of the central phraseological items in linguistic book reviews, and how it helps to determine the extent of the phraseological tendency of language.

The four steps of the PP model help the researcher to uncover the phraseological profile of a text or text type and address the following questions:

- What are the central phraseological items in a text or text collection? (Step 1: identification of phraseological items)
- How variable are these phraseological items? What types of internal variation do they allow? (Step 2: determination of item-internal variation)
- What functions do these phraseological items most commonly express? (Step 3: examination of the functions of identified items)
- How does the occurrence of the phraseological items relate to text structure? (Step 4: analysis of item distribution across texts)

The four steps provide important information about the text type under analysis and allow insights into the ways in which meanings are created in the discourse of a community or an individual (if, for instance, only texts produced by a particular author are examined). The outcome of these steps is a text-type specific inventory of phraseological items together with their variation, functions, and textual distribution. The PP model integrates some core features of current corpus linguistic practice and shows how Sinclair's (1996) search for units of meaning can be continued with more powerful software tools and new analytic techniques.[9] It also provides interesting insights into the extent of Sinclair's Idiom Principle and adds further supportive evidence on the phraseological tendency of language: words do not appear in isolation but "go together and make meanings by their combinations." (Sinclair 2004: 29).

In this article the PP model has been applied to an electronic collection of academic book reviews from the discipline of linguistics. Although only a selection of results from the book review corpus study have been discussed, it has hopefully become clear that the PP model enables views on the data that other book review studies have not been able to provide. Focussing on recurring phraseological items and their characteristics can shed light on the construction of meaning in academic book reviews.

The analyses reported on in this paper have also shown that a lot of important information about the co-selection and textual distribution of words and phrases has not yet been captured in linguistic analysis and description. Further studies exploring the patterned nature of language or, rather, well-defined subsets of it are still required. Findings based on applications of the PP model could have implications for the creation of text-type, discipline or genre specific reference works. The main implications I see, however, are of pedagogical nature. The PP model applied to a particular text type can help answer the question "What do learners need to know about the use of common phrases (in that particular text type)?" The present book review study has highlighted items that may be of use to novice academic

writers in the field of linguistics. Novice academic writers might also profit from knowing how common phrases can be modified, what functions they express, and what position in a text they tend to occur in. Mastery of the writing norms and conventions of a community may help learners (and novice writers in general) become accepted members of the community they wish to belong to.

Acknowledgements

I am grateful to the members of the University of Michigan Corpus Analysis Group (http://ctr.elicorpora.info/university-of-michigan-corpus-analysis-group), in particular Nick Ellis, Diane Larsen-Freeman, Matthew Brook O'Donnell and John Swales, for insightful comments and suggestions on a version of this paper I presented at a group meeting in August 2008. Matthew Brook O'Donnell's help and programming expertise was invaluable too in developing a system of tracking item occurrence across texts. I would also like to thank participants of the Corpus Linguistics 2009 conference in Liverpool for stimulating questions and positive feedback after my presentation on a previous version of the PP model, and an anonymous reviewer for constructive comments on the paper.

Notes

1. For more information on restricted languages and the related concept of "sublanguage", see Römer (Forthcoming).

2. Of course, if someone was interested in the linguistic characteristics of introduction sections only, it would be justified to work with samples and restrict the text selection to the initial part of each text.

3. Other software programs which can be used to extract n-gram lists are *AntConc* (Anthony 2007) and *WordSmith Tools* (Scott 2008).

4. Among other things, *Collocate* also extracts collocations (of different spans) around specified search words, e.g. 3-word combinations with the word *fact*. Such collocation extracts were done sporadically to complement the results of the n-gram analyses.

5. For a discussion of a related concept, "collocational frameworks", see Renouf and Sinclair (1991).

6. *ConcGram* also allows for fully automated searches for 2-, 3- and 4-word association patterns. These searches, however, produce very long output lists of tens of thousands of patterns which are hard to process manually.

7. See the detailed discussion of this topic by Evert (2004).

8. It should be noted that even p-frames (according to our definition) with an internal variable slot may be part of larger frames. Since we are working with different spans, these frames would also be captured in our analysis.

9. See my comments on Neo-Firthian/Sinclairian linguistic analyses in Römer (Forthcoming).

References

Anthony, L. 2007. *AntConc 3.2.1w (Windows)*. Tokyo, Japan: Waseda University.

Barlow, M. 2004a. *Collocate 1.0: Locating collocations and terminology*. Houston, TX: Athelstan.

Barlow, M. 2004b. *MonoConc Pro 2.2 (MP2.2)*. Houston, TX: Athelstan.

Biber, D. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

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

Biber, D. 2009b. A corpus-driven approach to formulaic language in English: Extending the construct of lexical bundle. In *Anglistentag 2008 Tübingen. Proceedings*, Reinfandt, C. & L. Eckstein (eds). Trier: Wissenschaftlicher Verlag Trier. 367–378.

Biber, D., Leech, G., Johansson, S., Conrad, S. & E. Finegan. 1999. *Longman grammar of spoken and written English*. London: Longman.

Cheng, W. 2008. Concgramming: a corpus-driven approach to learning the phraseology of discipline-specific texts. *CORELL: Computer Resources for Language Learning* 1: 22–35.

Cheng, W., C. Greaves & M. Warren. 2006. From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics* 11(4): 411–433.

Ellis, N. C. 2009. Optimizing the input: Frequency and sampling in usage-based and form-focussed learning. In *Handbook of language teaching*, Long, M. H. & C. Doughty (eds). Oxford: Blackwell. 139–158.

Ellis, N. C. & T. Cadierno. 2009. Constructing a second language: Introduction to the special section. *Annual Review of Cognitive Linguistics* 7: 113–140.

Ellis, N. C. & F. Ferreira-Junior. 2009. Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics* 7: 188–221.

Ellis, N. C., M. B. O'Donnell, U Römer, S. T. Gries & S. Wulff. 2009. Measuring the formulaicity of language. Paper Presented at the American Association of Applied Linguistics Annual Conference 2009, Denver, CO, 21–24 March.

Evert, S. 2004. *The statistics of word cooccurrences: Word pairs and collocations*. University of Stuttgart: Institut für maschinelle Sprachverarbeitung. Available at: http://elib.uni-stuttgart.de/opus/volltexte/2005/2371/.

Firth, J. R. 1956. Descriptive linguistics and the study of English. 1968. *Selected Papers of J. R. Firth 1952–59*, F. R. Palmer (ed). Bloomington: Indiana University Press. 96–113.

Fletcher, W. H. 2002–2007. *KfNgram*. Annapolis, MD: USNA.

Gea Valor, L. 2000. *A pragmatic approach to politeness and modality in the book review articles*. Valencia, Spain: SELL Monographs.

Granger, S. & F. Meunier (eds.). 2008. *Phraseology: An interdisciplinary perspective*. Amsterdam: John Benjamins.

Greaves, C. 2005. *ConcGram concordancer with ConcGram analysis*. HongKong: HKUST.

Greaves, C. 2009. *ConcGram 1.0. A phraseological search engine*. Amsterdam: John Benjamins.

Gries, S. T. 2008. Phraseology and linguistic theory: A brief survey. *Phraseology: An interdisciplinary perspective*, Granger, S. & F. Meunier (eds). Amsterdam: John Benjamins. 3–25.

Gries, S. T. 2009. Bigrams in registers, domains, and varieties: A bigram gravity approach to the homogeneity of corpora. Paper presented at Corpus Linguistics 2009, University of Liverpool, 22 July 2009.

Hoey, M. P. 2005. *Lexical Priming: A new theory of words and language*. London: Routledge.

Hoey, M. P. 2009. Corpus-driven approaches to grammar: The search for common ground. *Exploring the lexis-grammar interface* Römer, U. & R. Schulze (eds). Amsterdam: John Benjamins. 33–47.

Hoey, M. P. & M. B. O'Donnell. 2008. Lexicography, grammar, and textual position. *International Journal of Lexicography* 21(3): 293–30.

Hofland, K. & S. Johansson. 1982. *Word frequencies in British and American English.* Bergen: Norwegian Computing Centre for the Humanities / London: Longman.

Hunston, S. 2002. *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Hunston, S. & G. Francis. 2000. *Pattern Grammar. A corpus-driven approach to the lexical grammar of English*. Amsterdam: John Benjamins.

Hyland, K. 2000. *Disciplinary discourses. Social interactions in academic writing*. London: Longman.

Mahlberg, M. & M. B. O'Donnell. 2008. A fresh view of the structure of hard news stories. *Online proceedings of the 19th European Systemic Functional Linguistics conference and workshop, Saarbrücken, 23–25 July 2007*, Neumann, S. & E. Steiner (eds). Available at http://scidok.sulb.uni-saarland.de/volltexte/2008/1700/.

Meunier, F. & S. Granger. (eds). 2008. *Phraseology in foreign language learning and teaching*. Amsterdam: John Benjamins.

Motta-Roth, D. 1998. Discourse analysis and academic book reviews: A study of text and disciplinary cultures. *Genre studies in English for academic purposes*, Fortanet, I., S. Posteguillo, J. C. Palmer & J. F. Coll (eds). Castelló, Spain: University of Jaume I. 29–58.

Mukherjee, J. & S. T. Gries. 2009. Lexical gravity across varieties of English: An ICE-based study of speech and writing in Asian Englishes. Paper presented at ICAME 30, University of Lancaster, 31 May 2009.

O'Donnell, M. B. In preparation. The Adjusted Frequency List: Evaluating a method to produce cluster-sensitive frequency counts.

Partington, A. 1998. *Patterns and meanings. Using corpora for English language research and teaching.* Amsterdam: John Benjamins.

Renouf, A. & J. M. Sinclair. 1991. Collocational frameworks in English. *English corpus linguistics. Studies in Honor of Jan Svartvik*, Altenberg, B. & K. Aijmer (eds). London: Longman. 128–143.

Römer, U. 2005a. *Progressives, patterns, pedagogy. A corpus-driven approach to English progressive forms, functions, contexts and didactics*. Amsterdam: John Benjamins.

Römer, U. 2005b. 'This seems somewhat counterintuitive, though…' — Negative evaluation in linguistic book reviews by male and female authors. *Strategies in academic discourse,* Tognini Bonelli, E. & G. Del Lungo Camiciotti (eds). Amsterdam: John Benjamins. 97–115.

Römer, U. 2008. Identification impossible? A corpus approach to realisations of evaluative meaning in academic writing. *Functions of Language* 15(1): 115–130.

Römer, U. 2009. The inseparability of lexis and grammar: Corpus linguistic perspectives. *Annual Review of Cognitive Linguistics* 7: 141–163.

Römer, U. Forthcoming. Observations on the phraseology of academic writing: Local patterns — local meanings? *Chunks in the description of language. A tribute to John Sinclair*, Herbst, T., S. Schüller & P. Uhrig (eds). Berlin: Mouton de Gruyter.

Römer, U. & M. B. O'Donnell. 2009. Exploring the variation and distribution of academic phrase-frames in MICUSP. Presentation at Corpus Linguistics 2009, University of Liverpool, UK, July 2009.
Römer, U. & R. Schulze. (eds) 2008. *Patterns, meaningful units and specialized discourses* [*International Journal of Corpus Linguistics* 13(3), Special Issue]. Amsterdam: John Benjamins.
Römer, U. & R. Schulze. (eds) 2009. *Exploring the lexis-grammar interface*. Amsterdam: John Benjamins.
Scott, M. 2008. *WordSmith Tools 5.0*. Liverpool: Lexical Analysis Software.
Scott, M. & C. Tribble. 2006. *Textual patterns. Key words and corpus analysis in language education*. Amsterdam: John Benjamins.
Sinclair, J. M. 1987. The nature of the evidence. *Looking up: An account of the COBUILD project in lexical computing*, Sinclair, J. M. (ed). London: HarperCollins. 150–159.
Sinclair, J. M. 1991. *Corpus concordance collocation*. Oxford: Oxford University Press.
Sinclair, J. M. 1996. The search for units of meaning. *Textus* IX(1): 75–106
Sinclair, J. M. 2004. *Trust the text. Language, corpus and discourse.* London: Routledge.
Sinclair, J. M. 2005. The phrase, the whole phrase, and nothing but the phrase. Proceedings of the Phraseology 2005 conference, Cosme, C., C. Gouverneur, F. Meunier & M. Paquot (eds). Université catholique de Louvain: Louvain-la-Neuve. 19–22.
Sinclair, J. M. 2008a. The phrase, the whole phrase, and nothing but the phrase. *Phraseology: An interdisciplinary perspective*, Granger, S. & F. Meunier (eds). Amsterdam: John Benjamins. 407–410.
Sinclair, J. M. 2008b. Preface. *Phraseology: An interdisciplinary perspective*, Granger, S. & F. Meunier (eds). Amsterdam: John Benjamins. xv–xviii.
Stubbs, M. 2001. *Words and phrases. Corpus studies of lexical semantics*. Oxford: Blackwell Publishers.
Stubbs, M. 2007. Quantitative data on multi-word sequences in English. The case of the word 'world'. *Text, discourse and corpora*, Hoey, M., M. Mahlberg, M. Stubbs & W. Teubert (eds). London: Continuum. 163–189.
Suárez-Tejerina, L. 2005. Is evaluation structure-bound? An English-Spanish contrastive study of book reviews. *Strategies in academic discourse*, Tognini Bonelli, E. & G. Del Lungo Camiciotti (eds). Amsterdam: John Benjamins. 117–132.
Tognini Bonelli, E. 2001. *Corpus linguistics at work*. Amsterdam: John Benjamins.
Wulff, S. 2009. *Rethinking idiomaticity. A usage-based approach*. London: Continuum.
Zipf, G. K. 1935. *The psycho-biology of language: An introduction to dynamic philology*. Cambridge, MA: M.I.T. Press.

Author's address

Ute Römer
University of Michigan
English Language Institute
500 E. Washington St.
Ann Arbor, MI 48104
U. S. A.

uroemer@umich.edu