The inseparability of lexis and grammar

Corpus linguistic perspectives*

Ute Römer
University of Michigan

This paper focuses on the interface of lexis and grammar and provides corpus evidence for the inseparability of two areas that have traditionally been kept apart, both in language teaching and in linguistic analysis and description. The paper will first give an overview of a number of influential research strands and model-building attempts in this area (Pattern Grammar and Collostructional Analysis, among others) and then explore the use of a selected lexical-grammatical pattern, the introductory *it* pattern (e.g. *it is essential for EFL learners to come to grips with connotations*, attested example) in corpora of expert and apprentice academic writing.

Keywords: corpus linguistics, lexis-grammar inseparability, introductory *it* patterns, apprentice vs. expert academic writing, proficiency development

1. Introduction

If there is one major finding of modern (computer) corpus linguistic research over the past 40 years, it is probably that language is highly patterned. To a high degree, language is made up of fixed or semi-fixed units, and the co-selection of language items can be predicted on the basis of research findings in the areas of collocation and phraseology (see the seminal work of John Sinclair; Sinclair, 1991, 2004; Sinclair, Jones and Daley, 1970/2004; and the publications in Granger and Meunier, 2008). Corpus studies, based on large collections of authentic text from a range of different sources, have provided massive evidence for the interdependence of lexis and grammar (or vocabulary and syntax). They have demonstrated that two areas that have traditionally been kept apart, both in language pedagogy and in linguistic theory, are in fact inseparable. As Hoey and O'Donnell (2008, p. 293) put it, "[i]n the traditional view […], there is a grammar for every language and there is, quite separately, a lexicon." As we now know, thanks to researchers like Sinclair

Annual Review of Cognitive Linguistics 7 (2009), **140–162**. DOI 10.1075/arcl.7.06rom
ISSN 1572–0268 / E-ISSN 1572–0276

and Hoey, among many others, this grammar-lexicon dichotomy may hold true for sentences which have been invented in order to illustrate it, but it collapses when we consult real language data.

This paper sketches some influential research strands and model-building attempts at the lexis-grammar interface and summarises insights gained from corpus linguistics on aspects of lexis-grammar co-selection. It thereby aims to provide an overview of relevant research and central concepts in this area. Presenting a case study on the use of the introductory *it* pattern (also referred to as anticipatory *it* pattern; cf. Biber et al., 1999, p. 1019; Leech and Svartvik, 2002, pp. 219, 295–297; Quirk et al., 1985, pp. 1224, 1391–1392) in corpora of apprentice and expert academic writing, the paper will then explore and exemplify how lexis and grammar are interrelated. Moreover, the potential influence of language proficiency on lexical-grammatical selection will be investigated. The paper finishes with some summarizing and concluding thoughts.

2. Corpus research at the lexis-grammar interface: Major strands

While phraseology has been at the periphery of language analysis for most of the 20th and 21st century and a marginal aspect of study in most linguistic circles (cf. Ellis, 2008), a growing number of researchers in corpus linguistics now focus on phraseological items, patterns, constructions, or multi-word units (see, for example, the contributions in Granger and Meunier, 2008; Meunier and Granger, 2008; Römer and Schulze, 2008; Römer and Schulze, 2009; Schmitt 2004).

The following sections of the paper will focus on a selection of central approaches that integrate grammar and lexis. It will attempt to summarize their core claims and discuss how they have helped advance current linguistic (and applied linguistic) thinking. The six research strands or theories that have been selected are: John Sinclair's Idiom Principle, Susan Hunston and Gill Francis's Pattern Grammar, Michael Hoey's Lexical Priming, Douglas Biber et al.'s Lexical Bundles, Stefan Gries and Anatol Stefanowitsch's Collostructional Analysis, and Construction Grammar, in its modern Goldbergian version.[1]

2.1 The Idiom Principle

Let me start with a principle and related concepts put forward by John Sinclair in the late 1980s and early-mid 1990s. Sinclair, perhaps the most innovative and influential figure in modern computer corpus linguistics, was clearly a pioneer in data-rich language analysis, placing the study of meaning at centre stage. Through his work on lexical items and collocation, Sinclair paved the way for research at the

lexis-grammar interface to take place and helped bring phraseology back in the focus of attention. Sinclair (1987a, 1987b, 1991, 1996) put forward two conflicting principles to explain how meanings are created in text: the open-choice principle and the idiom principle. The open-choice principle (also referred to as the 'slot-and-filler' model) sees "language text as the result of a very large number of complex choices." (Sinclair, 1991, p. 109 and 1987a, p. 320) Grammars which assume that the slots in a sentence are more or less randomly filled by words (only making sure that the result is grammatical) are based on this principle.

However, since "words do not occur at random in a text, and […] the open-choice principle does not provide for substantial enough restraints for consecutive choices" (Sinclair, 1991, p. 110), there is the need for a second principle that accounts for further constraints: the idiom principle. According to the idiom principle, "a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments." (Sinclair, 1991, p. 110) The idiom principle refers to the "phraseological tendency" of language, i.e. the fact that words do not appear in isolation but "go together and make meanings by their combinations." (Sinclair, 2004, p. 29) This can be easily illustrated by the phrase *of course* (one of the examples Sinclair uses), which consists of two words but behaves in the same way as one-word adverbials like *sure*, *perhaps*, or *maybe*, and the components of which (*of* and *course*) are "not the preposition *of* that is found in grammar books" and "not the countable noun that dictionaries mention" (Sinclair, 1991, p. 111) but take on meaning in the phrase (most recently, Sinclair referred to this idea by the term "meaning shift unit", Sinclair, 2007, personal communication). Another related Sinclairian notion is that of "lexical grammar", which is "an attempt to build together a grammar and lexis on an equal basis" (Sinclair, 2004, p. 164), hence aiming for a true integration of structural (or syntactic) patterns and vocabulary patterns.

2.2 Pattern Grammar

Firmly based in a Birmingham (and Sinclairian) corpus-driven tradition, Susan Hunston and Gill Francis's Pattern Grammar echoes the notion of lexical grammar (the term appears in the subtitle of Hunston and Francis's 2000 book *Pattern Grammar*), and provides a further development of some of Sinclair's ideas and theories. Pattern Grammar is "an approach to lexis and grammar based on the concept of phraseology and of language patterning arising from work on large corpora." (Hunston and Francis, 2000, cover blurb) Patterns are phraseological items, i.e. neither single words nor empty grammatical structures (the slots of which are filled with words) but results of a synthesis of the two. Patterns show how words

are typically associated with each other and how they form meaningful units. The 'V *over* N' pattern (a verb followed by *over*, followed by a noun; Hunston and Francis, 2000, p. 43), for example, indicates that it is common for the preposition *over* to be immediately preceded by a verb and, in such cases, to occur right before a noun or noun group. Together, a verb, *over*, and a noun form a unit of meaning, as in *fight over Europe*, *grieved over her*, and *triumph over Russia* (examples taken from Hunston and Francis, 2000, p. 43, 44).

By discussing how patterns are formed to express meanings, the authors provide evidence for Sinclair's (1991, p. 65) observation that "there is a strong tendency for sense and syntax to be associated". They state and exemplify that "the different senses of words will tend to be distinguished by different patterns, and secondly, that particular patterns will tend to be associated with lexical items that have particular meanings." (Hunston and Francis, 2000, p. 83) This means that patterns can be used to help us distinguish the different meanings of a polysemous word. It also means that we can derive aspects of meaning of a word in a pattern from the meanings of other semantically related words that occur in the same slot in the same pattern. Hunston and Francis illustrate this sense and syntax association by means of the verb *reflect* and three of its meanings (Hunston and Francis, 2000, p. 255–256). They observe that each of the three meanings (light-related *reflect*, mirror-related *reflect*, thinking-related *reflect*) "typically occurs in a particular phraseology, that is, collocating with different types of noun or pronoun […] and with a different complementational pattern" (Hunston and Francis, 2000, p. 255), e.g. V + N (*to reflect light*), *be* V-*ed* (*sth. is reflected in a mirror*), or V + Prep (*I reflected on sth.*).

I will return to some of the Pattern Grammar principles in the discussion of introductory *it* pattern subtypes in Section 3. This will look at the relations between the adjectives found in and meanings expressed by these pattern types.

2.3 Lexical Priming

Another scholar in the British contextualist tradition who has been influenced by and further develops ideas of John Sinclair is Michael Hoey. In his 2005 book entitled *Lexical Priming* he proposes "a new theory of the lexicon, which amounts to a new theory of language" and which "contextualises theoretically and psychologically Sinclair's insights about the lexicon" (Hoey, 2005, p. 1 and p. 158).[2] Hoey not only aims at integrating vocabulary and syntax, but puts lexis at centre stage. He states: "[t]he theory reverses the roles of lexis and grammar, arguing that lexis in complexly and systematically structured and that grammar is an outcome of this lexical structure." (Hoey, 2005, p. 1) He also goes futher than Sinclair by making psycholinguistic claims.

Central to Hoey's theory is the observation that "[e]very word is primed for use in discourse as a result of the cumulative effects of an individual's encounters with the word." (Hoey, 2005, p. 13; see also Hoey, 2004, p. 386, and Hoey, 2009) In other words, as we encounter words in spoken and written discourse and use them ourselves, we automatically pick up their usage patterns and learn in which language structures, textual positions, or text types they typically appear. In this process, existing primings can either be reinforced or weakened (see Hoey, 2005, p. 9). As a result, our knowledge about a word is, according to Hoey, entirely dependent on our experiences with it (i.e. on how we have seen/heard it being used and how we have used it ourselves). This implies that, among other things, priming effects are register-specific and that large collections of different text types have to be studied separately so that we discover how for example newspaper texts prime us differently in using a word than TV sitcoms do.

Like Sinclair's and Hunston and Francis's, Hoey's view of language stands in contrast with traditional views which treat lexis and grammar separately. Hoey promotes an approach which starts from vocabulary items and then looks at their favoured associations and usage patterns. In this context, he considers it important that linguistic theories should focus on what is *natural* in a language and need to acknowledge the pervasiveness of collocation. Grammar then is "the product of the accumulation of all the lexical primings of an individual's lifetime" (Hoey, 2005, p. 159), the outcome of combining collocational primings in such a way that they form a system.

2.4 Lexical Bundles

Moving on from the British contextualist tradition approaches to North American corpus linguistic thinking, another important strand of research at the interface of lexis and grammar is Douglas Biber's Lexical Bundles work. Biber and his colleagues take an integrative stance on lexis and grammar by looking at repeated contiguous word combinations, or multi-word units (MWUs), sequences of three or more words, across spoken and written registers. Lexical bundles are defined as "recurrent expressions, regardless of their idiomaticity, and regardless of their structural status." (Biber et al., 1999, p. 990) To be classified as a lexical bundle, an MWU has to (a) occur frequently in a register, e.g. 10 times per one million words,[3] and (b) occur in multiple texts in this register. The dispersion measure is considered so to make sure that the repeated occurrence of a word combination is not due to speaker/writer idiosyncrasies. Examples of lexical bundles found by Biber and colleagues to be frequent in academic writing include *in order to*, *one of the*, *as a result of*, *it is possible to*, and *on the other hand* (Biber et al., 1999, p. 994).

A particularly interesting aspect of lexical bundles is that they, more often than not, cross the boundaries of traditional grammatical categories such as noun phrases or prepositional phrases. As Biber, Conrad and Cortes (2004, p. 377) point out, "most lexical bundles do not represent a complete structural unit." Instead, they very often, as examples like *as a result of* or *it is possible to* demonstrate, "bridge two structural units: they begin at a clause or phrase boundary, but the last words of the bundle are the first elements of a second structural unit." (Biber, Conrad and Cortes, 2004, p. 377) What Biber et al.'s research indicates is that "lexical bundles are a fundamentally different kind of linguistic construct from productive grammatical constructions" (Biber, Conrad and Cortes, 2004, p. 399) and clearly an important kind of construct that deserves further attention. So far the lexical bundle approach has mainly been used to account for register and text-type differences, to identify meaningful and organisational units in texts, and to study differences between native- and non-native speaker output (see, for example, Biber, 2006; Cortes, 2004; de Cock, 1998; Hyland, 2008b), but there is certainly scope for further applications, both in linguistics and in language teaching, e.g. in studying historical or regional language variation, or in cross-linguistic analyses.

2.5 Collostructional Analysis

A more recent but already influential approach that studies the integration of lexis and grammar and aims to reconcile corpus and cognitive linguistics is Collostructional Analysis, a family of analytic techniques initially developed and put forward by Stefan Gries and Anatol Stefanowitsch. Collostructional Analysis is "an extension of collocational analysis specifically geared to investigating the interaction of lexemes and the grammatical structures associated with them." (Stefanowitsch and Gries, 2003, p. 209) It measures the association strength between smaller constructions (morphemes or words) and the more complex constructions they occur in. The central question here is "Are there significant associations between words and grammatical structure at all levels of abstractness?" (Stefanowitsch and Gries, 2003, p. 211) That means that Collostructional Analysis investigates which words occur more (or less) frequently than expected in a particular construction, e.g. which verbs are associated with ditransitives in English (another example would be the occurrence of verbs in the '*go* (*and*) V' construction; see Wulff, 2006).

A subtype of Collostructional Analysis is distinctive collexeme analysis (DCA) which is "specifically geared to investigating pairs of semantically similar grammatical constructions and the lexemes that occur in them" (Gries and Stefanowitsch, 2004, p. 97). DCA serves to uncover the extent to which lexical items are attracted (or repelled) by a particular grammatical structure and hence demonstrates in what ways (and how closely) lexis and grammar interact with each other. One of

the several concrete examples that Gries and Stefanowitsch (2004) discuss in this context is the alternation between the *s*-genitive and the *of*-genitive in English. Based on data from ICE-GB (the British component of the International Corpus of English), the authors identify words that most distinctively occur as heads and modifiers in both constructions (e.g. *friend*, *mother*, and *father* are highly distinctive head nouns in the *s*-genitive, as in *Elena's friend*; Gries and Stefanowitsch, 2004, p. 117). Wulff and Gries (this issue) provide further information on Collostructional Analysis and its use in studying constructions in L2 production.

2.6 Construction Grammar

A final relevant approach I would like to summarize here is modern Construction Grammar as put forward by Adele Goldberg (see Goldberg, 1995 and 2006). Construction Grammar, "developed by renegades from an American Chomskyan cognitive tradition" (Stubbs, 2009, p. 27), is not actually a strand in corpus linguistics and uses corpora much less than the approaches described in Sections 2.1 to 2.5, but it represents usage-based research at the interface of lexis and grammar.

Constructions, defined as "conventionalized pairings of form and function" (Goldberg, 2006, p. 3) and stored as units in the brain, exist on all levels of grammatical analysis and cover "morphemes and words, idioms, partially lexically filled and fully general phrasal patterns." (Goldberg, 2006, p. 5) Hence the prefix *un-*, the adjective *happy*, the compound *lovesick*, the idiom *make hay while the sun shines*, and the progressive (form of BE + present participle) are all constructions. In its entirety, the extended mental lexicon renders a distinction between lexis and grammar obsolete (see Goldberg, 2006, p. 18). Phenomena that have been looked at from a Construction Grammar perspective include argument structure constructions (including distransitives; Goldberg, 1995, 2006), resultative constructions (Goldberg and Jackendoff, 2004), and future constructions (Hilpert, 2008).

A key aspect about constructions is that they are (similar to patterns in the Pattern Grammar framework) "highly valuable both in predicting meaning, given the form, and in predicting form, given the message to be conveyed." (Goldberg, 2006, p. 228) This implies that constructions play a crucial role in facilitating communication: Based on our previously acquired constructional knowledge (this reminds us of Hoey's priming), they enable us "to understand and produce utterances" (Goldberg, 2006, p. 228). Construction Grammar claims to be usage-based and places some emphasis on frequency as an indicator for the existence of constructions. Even semantically transparent and functionally predictable patterns qualify as constructions, provided that they are sufficiently frequent. In Goldberg's words, usage-based Construction Grammar implies that "facts about the actual

use of linguistic expressions such as frequencies and individual patterns that are fully compositional are recorded alongside more traditional linguistic generalizations." (Goldberg, 2006, p. 45) Despite the emphasis put on frequencies, however, Goldberg's 1995 and 2006 books contain only little corpus data but mainly constructed examples or examples borrowed from earlier studies.

2.7 Shared features of the different strands

Different as the approaches we have just described may appear, they have a number of things in common. The first observation common to the Idiom Principle, Pattern Grammar, Lexical Priming, Lexical Bundles, Collostructional Analysis, and, though to a lesser extent, Construction Grammar is that the study of language is empirical and based on large amounts of naturally occurring text. Frequency of occurrence (and co-occurrence) of language items is crucial, and corpora and corpus tools (software packages, computer scripts or online search interfaces) are used to identify which items are common in which contexts and in which types of discourse. For most researchers mentioned in the above sections, data and observation come first, and theory comes second. Hoey's Lexical Priming theory and Hunston and Francis's Pattern Grammar, for instance, provide explanations for what is observed in the language. Biber and colleagues look at "descriptive facts that require explanation" (Biber, Conrad and Cortes, 2004, p. 400). While these approaches derive theoretical findings inductively from the data, Construction Grammar is different in that it places theory before observation (see also Hunston, 2008, p. 292).

The core observations as to the interrelatedness of vocabulary and syntax and the conclusions the featured approaches arrive at, however, are largely similar which, according to Hunston (2008, p. 292) "would tend to increase confidence" in all of them. They all find that form and meaning are inseparable and that the unit of meaning in language is not the word in isolation but a construction or phrasal unit (at different levels of complexity). The pervasiveness of co-selection features and collocations is emphasized in all strands; differences here are mainly terminological. Finally, the most important shared observation which connects all described strands is that it is impossible to divorce lexical items and grammatical constructions and that phraseological items should play a more central role in linguistic theory and description. The following case study will deal with a selected phraseological item and testify the inseparability of lexis and grammar.

3. A case study: The introductory *it* pattern and language proficiency development

Let us now turn from the theoretical foundations of current data-intensive phraseological research to a case study at the interface of lexis and grammar which also incorporates a language learning dimension. In this case study I will examine the lexical realisations of the so-called introductory *it* pattern (as realised in *it is essential for EFL learners to come to grips with connotations*, attested example) and analyse potential connections between this pattern (and its subpatterns) and language proficiency development — or, more precisely, the development of native and non-native speakers' academic writing proficiency. In the context of language learning, the introductory *it* pattern constitutes a particularly interesting phenomenon because it is known to cause problems for EFL learners (cf. Hewings and Hewings, 2002, p. 368). These problems are worth addressing, especially from an EAP teaching perspective, given that introductory *it* patterns are very common in academic writing across disciplines (see Groom, 2005; Hyland, 2008a; Oakey, 2002).

3.1 Data

The case study is based on data derived from four corpora of apprentice and expert academic writing, i.e. writing produced in academic settings or in a university context by different groups of students or academics. Following Scott and Tribble (2006, p. 133), apprentice texts are understood to be "unpublished pieces of writing that have been written in educational or training settings", whereas expert texts are pieces of writing that have been published. The corpora used are listed in Table 1 together with their size and a brief description of the type of data they capture.

The group of 'apprentice' writers (captured in GICLE, CHALC, and MICUSP_HS) covers both non-native and native speaker writers on different levels of proficiency. GICLE, the German part of the International Corpus of Learner English (Granger et al., 2002), consists of undergraduate student argumentative essays, i.e. writing samples by upper-intermediate learners (L1 German). Covered in CHALC, the Cologne-Hanover Advanced Learner Corpus (Römer, 2007), are linguistics and literary studies essays and term papers written mainly by final year undergraduates and first year graduates who can be classified as advanced learners (L1 German). MICUSP_HS, a subsection of the Michigan Corpus of Upper-level Student Papers (under compilation at the University of Michigan English Language Institute; see http://micusp.elicorpora.info) consisting of papers from the Humanities and Social Sciences, is a collection of writing samples by mainly (in this subset around 75%) American English native speaker graduate and final year undergraduate students who were enrolled in degree programmes at the University of Michigan at

Table 1. Corpora used in the study

Name	Description	Size
GICLE	German component of the International Corpus of Learner English; 450 argumentative essays by undergraduate students	~ 234,000 words
CHALC	Cologne-Hanover Advanced Learner Corpus; 45 linguistics/literary studies essays and term papers by upper-level students (final year undergraduates and first year graduates)	~ 200,000 words
MICUSP_HS	Humanities and Social Sciences subsection of the Michigan Corpus of Upper-level Student Papers (http://micusp.elicorpora.info); 162 A-grade writing samples from final year undergraduates and first through final year graduate students (~ 75% NS of AmE) in Linguistics, Philosophy, Psychology, and Sociology	~ 470,000 words
Hyland_HS	Humanities and Social Sciences subsection of the Hyland Corpus (Hyland 1998); 90 published research articles from Linguistics, Philosophy and Social Sciences	~ 611,000 words

Ann Arbor. Since papers by second and third year graduate students are included here, MICUSP_HS writers can on average be considered more advanced in terms of their academic writing proficiency, compared to CHALC writers. In addition to these three corpora of apprentice production data, I consulted a corpus of expert academic writing: Hyland_HS, the Humanities and Social Sciences subsection of the Hyland Corpus (Hyland, 1998). Hyland_HS consists of 90 published research articles (30 each) from Linguistics, Philosophy, and Social Sciences, and nicely matches MICUSP_HS in terms of its disciplinary coverage.

As should become apparent from the corpora descriptions, there are at least three potentially influential variables that may affect our results and that we will need to take into account in interpreting our findings: (i) nativeness (GICLE and CHALC vs. MICUSP_HS and Hyland_HS),[4] (ii) general language proficiency level (GICLE vs. CHALC; GICLE vs. MICUSP_HS/Hyland_HS; CHALC vs. MICUSP_HS/Hyland_HS), and (iii) expertise in academic writing or academic writing proficiency, represented by the number of years of higher/university education (increasing from GICLE via CHALC via MICUSP_HS to Hyland_HS). I will refer back to these variables when I discuss the distribution, types and functions of introductory *it* patterns across corpora in Section 3.3 below.

3.2 Method

The identification of the introductory *it* pattern as an intriguing lexical-grammatical phenomenon of academic discourse was part of a larger scale phraseological

exploration of apprentice and expert academic writing. In this exploration, two new-generation corpus tools that can be classified as 'phraseological search engines' were used. The tools used were Michael Barlow's *Collocate* (Barlow, 2004) and William Fletcher's *kfNgram* (Fletcher, 2002–2007). *Collocate* extracts lists of n-grams of different lengths (i.e. combinations of n words) and collocations (word clusters) with specified search words from a corpus. Examples of n-grams (span = 4; cf. Biber et al.'s lexical bundles) that frequently occur in academic writing corpora are *at the end of*, *at the same time*, or *on the other hand*. The 4-gram *at the end of* also occurs in a 4-word Collocate collocations list if *end* is used as search term and the span set to 4. *KfNgram* (like *Collocate*) also generates lists of n-grams from a corpus, and in addition to that, lists of so-called 'phrase-frames' (short '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 is at work (i.e. how fixed language units are or how much they allow for variation).

In our search for phraseological items in expert academic writing (the considered 'target' text type for our apprentice writers) the first analytic step was the extraction of n-grams and p-frames of different lengths from Hyland_HS. The resulting *Collocate* and *kfNgram* output lists contained a large number of *it* patterns, e.g. *it is not*, *it is a*, *it is rational for*, and *it is true that*. Particularly frequent among the 4-word items were the p-frame *it is * to* and its realisations *it is important to*, *it is possible to*, and *it is difficult to*. These common patterns around *it* were then taken as a starting point to investigate lexical-grammatical choices in academic writing on different levels of proficiency. Concordance searches for *it is* in the four selected corpora resulted in 4,000 hits altogether — 759 hits in GICLE (i.e. 3,243 per million words), 467 in CHALC (2,335 pmw), 1,147 in MICUSP_HS (2,440 pmw), and 1,627 in Hyland_HS (2,662 pmw).[5]

Next was a manual analysis and filtering of the 4,000 retrieved concordance lines. Lines that did not contain a form of the '*it is* (ADV) ADJ' pattern were deleted from the database and hence excluded from the subsequent formal and functional analyses. That means the remaining concordance lines all contained *it is* followed by and adverbial (optional), followed by an adjective (obligatory). There are of course also variants of the introductory *it* pattern where the linking verb (mainly BE, but also sometimes SEEM or APPEAR) is followed by the past participle of a verb or by a noun phrase, as in *it is acknowledged that two cases are insufficient* or *it is a beginning of one and an end of another* (both examples from Hyland_HS). These pattern types were, however, much less frequent in our data than the *it is* (ADV) ADJ pattern and would have further complicated the picture of grammatical patterns and lexical choices.

After the manual filtering process, the following sets of (altogether 1,485) concordance lines remained: 228 from GICLE (974 pmw), 156 from CHALC (780 pmw), 512 from MICUSP_HS (1,093 pmw), and 589 from Hyland_HS (964 pmw). These concordance lines were then classified according to pattern subtypes. The two major types were '*it is* (ADV) ADJ *to*-infinitive' (e.g. *it is necessary to look briefly at the concerns of this type of genre*) and '*it is* (ADV) ADJ *that*-clause' (e.g. *it is clear that Aristotle intended something much broader),* while a small number of remaining examples (mainly of the '*it is* (ADV) ADJ *wh*-word' type) made up a third group of subpatterns labelled 'other'. This subtype classification and the initial frequency counts based on the four corpora then led to the first set of results. In the analysis of the datasets from GICLE, CHALC, MICUSP_HS, and Hyland_HS, I also focussed on a functional classification of all examples (by subpattern), the distribution of adjectives across patterns, and the proportion and type of adjective modification. The results of these investigations will be discussed in turn below.

3.3 Findings

If we look at the overall (absolute normalised) frequencies of introductory *it* patterns (all subtypes taken together) across corpora (see Figure 1), we do not see any linear trends that would mirror the development of academic writing proficiency (we could have assumed that the use of such a typical pattern of academic writing increases with increasing proficiency), but instead find deviations from Hyland_HS (the expert/target norm) both for CHALC and MICUSP_HS, while the overall frequency of introductory *it* in GICLE is very similar to that in Hyland_HS. This finding may seem surprising, but we have to consider that we are here in fact dealing with *a set* of different patterns that may behave differently. So there may well be concurrent trends involved that balance each other out.

To see if this is the case, let us turn to Figure 2 which gives a graphic representation of the proportions of the three attested subpatterns in the four corpora. Indeed, we observe different trends for the two frequent subpatterns, *it is* (ADV) ADJ *to*-infinitive and *it is* (ADV) ADJ *that*-clause. The *it is* (ADV) ADJ *to*-infinitive pattern is most frequent in all four corpora but its frequency is lower in MICUSP_HS than in CHALC and lower in CHALC than in GICLE (and roughly on the same level in Hyland_HS and CHALC). On the other hand, the proportions of the *it is* (ADV) ADJ *that*-clause pattern are lowest in GICLE, higher in CHALC and even higher in MICUSP_HS. If we assume that MICUSP_HS writers are more experienced and proficient in academic writing than CHALC writers, who in turn are more proficient than GICLE writers, we can say that the observed proportions of the *it is* (ADV) ADJ *that*-clause pattern increase in order of increasing academic writing proficiency (from GICLE via CHALC to MICUSP_HS); Hyland_HS is

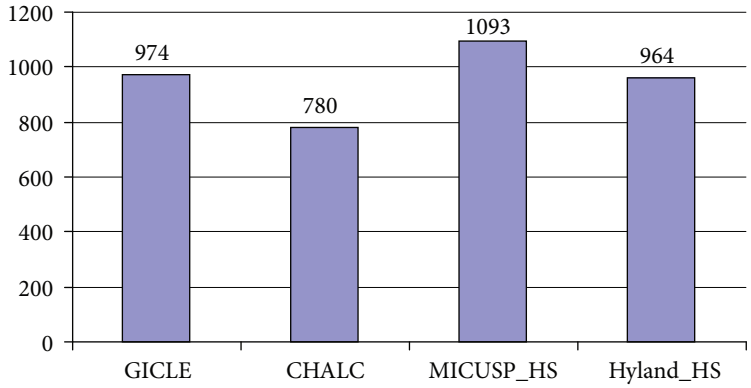
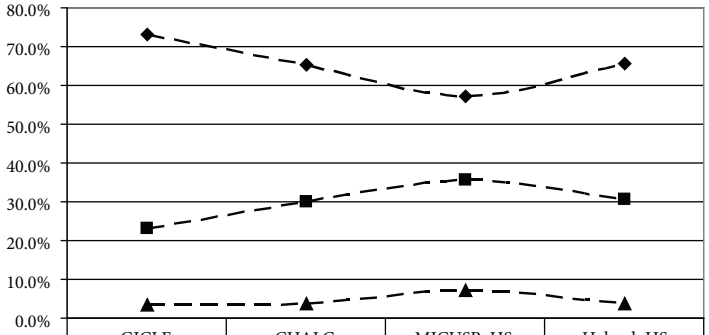


Figure 1. Frequencies (per million words) of introductory *it* patterns across corpora



	GICLE	CHALC	MICUSP_HS	Hyland_HS
it is (ADV) ADJ to-infinitive	73.2%	65.4%	57.2%	65.7%
it is (ADV) ADJ that-clause	23.2%	30.1%	35.6%	30.7%
it is (ADV) ADJ + other	3.5%	3.8%	7.2%	3.6%

Figure 2. Relative frequencies of introductory *it* subpatterns across corpora (corpora displayed in order of assumed proficiency development)

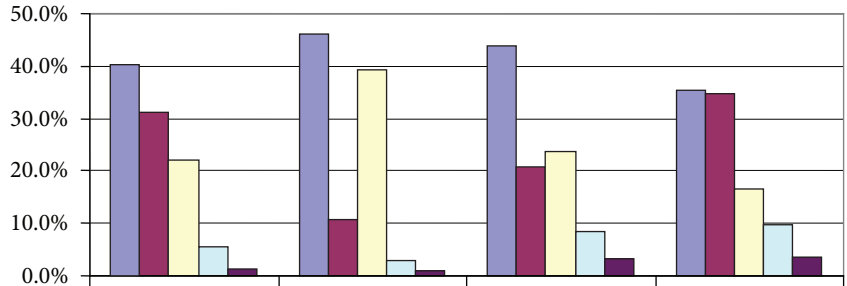
again very similar to CHALC. As we can see in Figure 2, GICLE writers exhibit a strong preference for the *it is* (ADV) ADJ *to*-infinitive pattern while the *it is* (ADV) ADJ *that*-clause pattern is comparatively underused by our German upper-intermediate learners. A possible explanation for this may lie in the learners's L1: Frequent parallel constructions in German (*Es ist* (ADV) ADJ *zu*-infinitive and *Das ist* (ADV) ADJ *zu*-infinitive) can be assumed to boost the *it is* (ADV) ADJ *to*-infinitive pattern in the German learner production data.

After these initial frequency counts, I moved on to the actual focus of the study and examined the relationship between the identified (sub)patterns and lexical choices in the adverbial (ADV) and adjective (ADJ) slots in the patterns. Part of this examination was a functional classification of each of the 1,485 introductory *it* examples from the four corpora which highlighted some interesting connections

between patterns and meanings — a recurrent theme in all of the research traditions described in Section 2 of this article. The functional classification was based on Groom's (2005) corpus study on introductory *it* patterns in research articles and book reviews in History and Literary Criticism. Following Francis, Hunston and Manning's (1998) *Grammar Patterns* model in which sets of semantically-related words that occur in the same slot in the same pattern (in this case adjectives in the introductory *it* pattern) are grouped under functional labels, Groom annotates his corpus data according to six functions: 'adequacy', 'desirability', 'difficulty', 'expectation', 'importance', and 'validity'.

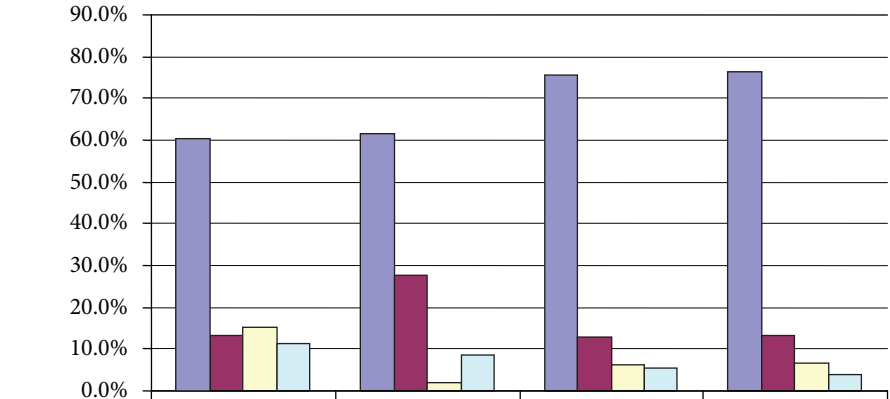
I was able to identify five out of these six functions (all except for 'adequacy') in my datasets and found that they are all strongly adjective- and pattern-related. Figures 3 and 4 illustrate the distributions of the five functions of the *it is* (ADV) ADJ *to*-infinitive pattern (Figure 3) and the *it is* (ADV) ADJ *that*-clause pattern (Figure 4) across corpora. Of the five identified functions, 'difficulty' is the most frequent one when it comes to the *to*-infinitive pattern but is never expressed by the *that*-clause pattern where 'validity' clearly tops the list of functions ('validity' comes last in Figure 3 and is only very rarely expressed by means of the *to*-infinitive pattern).

If we now look at Figure 3 and consider how our three apprentice writing corpora compare to Hyland_HS, we see that the picture is rather complex. 'Difficulty' is the most frequent function expressed by the *to*-infinitive pattern in all four corpora, but the relative frequencies are higher in GICLE, CHALC, and MICUSP_HS than in Hyland_HS. The proportions for 'desirability' are similar in GICLE and



	GICLE	CHALC	MICUSP_HS	Hyland_HS
difficulty	40.1%	46.0%	43.9%	35.4%
desirability	31.1%	10.8%	20.7%	34.9%
importance	22.2%	39.2%	23.8%	16.5%
expectation	5.4%	2.9%	8.5%	9.6%
validity	1.2%	1.0%	3.1%	3.6%

Figure 3. Functions of the *it is* (ADV) ADJ *to*-infinitive pattern across corpora



	GICLE	CHALC	MICUSP_HS	Hyland_HS
validity	60.4%	61.7%	75.7%	76.2%
expectation	13.2%	27.7%	12.7%	13.3%
importance	15.1%	2.1%	6.1%	6.6%
desirability	11.3%	8.5%	5.5%	3.9%

Figure 4. Functions of the *it is* (ADV) ADJ *that*-clause pattern across corpora

Hyland_HS (31.1% and 34.9%) but much lower in CHALC (10.85) and MICUSP_HS (20.7%). 'Importance' is much more often expressed in the three apprentice writing corpora (especially in CHALC) than in published/expert writing. Finally, while 'expectation' and 'validity' are rather infrequent functions in all four corpora, they are both somewhat more common in MICUSP_HS and Hyland_HS than in the two collections of German learner writing.

For the *it is* (ADV) ADJ *that*-clause pattern, the distribution of functions across corpora (see Figure 4) appears somewhat less complex than for the *it is* (ADV) ADJ *to*-infinitive pattern. The functional profile for MICUSP_HS is almost identical to Hyland_HS (which may indicate that our upper-level predominantly native-speaker student writers can handle the pattern as well as our expert writers), whereas both GICLE and CHALC writers deviate from this expert profile and express 'validity' less frequently but other functions ('expectation' in CHALC and 'importance' in GICLE) more frequently than MICUSP_HS and Hyland_HS writers. Among the two learner corpora distributions, there are similarities (concerning 'validity' and 'desirability') but also differences (concerning 'expectation' and 'importance'). In order to make sense of the deviations in terms of the functions encoded by means of both introductory *it* pattern by apprentice writers as compared to expert writers, I zoomed in on lexical choices and took a closer look at the adjectives used in all four corpora to express the five functions, again separately for the two dominant patterns, *it is* (ADV) ADJ *to*-infinitive and *it is* (ADV) ADJ *that*-clause.

Tables 2 and 3 list, in order of relative frequency, the top-10 adjectives found in both types of introductory *it* patterns across corpora. The first thing we note in Table 2 is a considerable cross-corpora overlap regarding adjective types (adjectives in small caps). *Possible*, *important*, *difficult*, *necessary*, *impossible*, and *easy* appear in the top-10 lists in all four corpora, however with rather different relative frequencies. In CHALC, for example, with 23.5% the two by far most frequently used adjectives in the *it is* (ADV) ADJ *to*-infinitive pattern are *important* and *possible*, which explains the high proportions of 'difficulty' and 'importance' we observed for this corpus in Figure 3. Still looking at Table 2, we also see that some adjectives only occur among the top-10 in one or two of the corpora (this can either mean that they do not occur in the other corpora at all or that they do occur but are very infrequent): *rational* and *interesting* only appear in the Hyland_HS list (the high frequency of *rational* explains the height of the 'desirability' bar in Figure 3), *reasonable* ('desirability') is only found in Hyland_HS and MICUSP_HS, and *better* (also expressing 'desirability') is only part of the top-10 lists for GICLE and CHALC.

Adjective type overlaps and deviating relative frequencies (e.g. for *true*, *clear*, *possible*, and *likely*) can also be found in Table 3, which lists the most common adjectives in the four corpora for the *it is* (ADV) ADJ *that*-clause pattern. *Possible* ('validity') is comparatively common in MICUSP_HS (with 25.9%), while the top choice of GICLE writers is *true* (34%; also belongs to the 'validity' group). Again, some adjectives are only shared by two or three corpora, e.g. *interesting*, which is a common choice in MICUSP_HS and (especially) CHALC — which helps to explain the comparatively high proportion found for the function 'expectation' in this corpus (see Figure 4), or *obvious* which is frequent in this pattern in Hyland_HS, MICUSP_HS, and CHALC.

Table 2. Top-10 adjectives in the *it is* (ADV) ADJ *to*-infinitive pattern across corpora

	GICLE	CHALC	MICUSP_HS	Hyland_HS
1	IMPORTANT 13.8%	IMPORTANT 23.5%	IMPORTANT 14.9%	rational 14.7%
2	DIFFICULT 11.4%	POSSIBLE 23.5%	POSSIBLE 12.9%	POSSIBLE 14.2%
3	POSSIBLE 8.4%	NECESSARY 11.8%	DIFFICULT 11.6%	IMPORTANT 9.0%
4	NECESSARY 7.2%	DIFFICULT 10.8%	IMPOSSIBLE 7.8%	DIFFICULT 6.7%
5	hard 6.6%	IMPOSSIBLE 5.9%	NECESSARY 5.8%	NECESSARY 5.4%
6	IMPOSSIBLE 6.0%	easier 2.9%	EASY 4.8%	IMPOSSIBLE 4.9%
7	better 4.8%	useful 2.9%	hard 3.4%	EASY 3.9%
8	easier 4.8%	better 1.0%	reasonable 3.1%	hard 3.4%
9	EASY 2.4%	EASY 1.0%	easier 2.4%	reasonable 2.8%
10	good 2.4%	effectual 1.0%	wrong 2.0%	interesting 2.3%

Table 3. Top-10 adjectives in the *it is* (ADV) ADJ *that*-clause pattern across corpora

	GICLE	CHALC	MICUSP_HS	Hyland_HS
1	TRUE 34.0%	TRUE 14.9%	POSSIBLE 25.9%	TRUE 18.2%
2	POSSIBLE 7.5%	POSSIBLE 12.8%	CLEAR 14.2%	CLEAR 14.9%
3	important 5.7%	interesting 10.6%	LIKELY 9.3%	POSSIBLE 11.6%
4	LIKELY 5.7%	obvious 10.6%	TRUE 7.7%	obvious 7.2%
5	necessary 3.8%	CLEAR 6.4%	SURPRISING 5.5%	SURPRISING 5.0%
6	SURPRISING 3.8%	SURPRISING 6.4%	important 2.7%	evident 3.9%
7	astonishing 1.9%	funny 4.3%	conceivable 2.2%	LIKELY 3.9%
8	CLEAR 1.9%	impossible 4.3%	interesting 2.2%	apparent 2.8%
9	common 1.9%	LIKELY 4.3%	obvious 2.2%	arguable 2.2%
10	dangerous 1.9%	natural 4.3%	apparent 1.6%	significant 2.2%

What I find particularly interesting in Tables 2 and 3 and further down on the adjective frequency lists, however, is the occurrence of a number of what I would call 'unexpected' or 'extreme' adjectives in the GICLE and (though to a lesser extent) in the CHALC columns. I am here referring to types such as (in alphabetical order) *amazing*, *astonishing*, *bad*, *dangerous*, *fascinating*, *frightful*, *funny*, *irresponsible*, *ridiculous*, *striking*, *scarying* [sic], *stupid*, *unbelievable*, *unfair*, *wonderful*, or *worst*. Unlike Hyland_HS and MICUSP_HS writers, the less experienced CHALC and in particular GICLE writers often use introductory *it* patterns to express strong emotions and personal opinions and get more involved than Hyland_HS and MICUSP_HS writers when they use introductory *it* as a means of evaluation. Not only do many of the adjectives found in the two learner corpora appear more emotional, they also tend to be adjectives that are more characteristic of speech than of academic writing (confirmed by cross-checks in the spoken and academic subsections of the Corpus of Contemporary American English, COCA; http://www.americancorpus.org/).

More evidence for an 'extreme' and speech-like tendency of GICLE (though not of CHALC) writers can be found in Figure 5 which illustrates to what extent adjectives (in both pattern subtypes) are modified by boosters, such as *very*, *certainly*, or *extremely*. With 20.6% the proportion of boosters is highest in the GICLE data which supports the above observation that our least advanced apprentice writers express subjectivity (or emotions) much stronger than higher-level apprentice and expert academic writers. In part this finding can be explained by the text type captured in GICLE (argumentative essays) which encourages the expression of a personal perspective on the issues discussed. The degree of emphatic modification and the number and variety of the adjectives used in GICLE are, however,

still remarkable, and the text type explanation does not work for CHALC which consists of linguistics and literary criticism essays and term papers but where we still find some of the 'unexpected' adjectives listed above. It is also worth mentioning in this context that Hewings and Hewings in their study on introductory *it* in student and published academic writing find that students make "certain inappropriate adjective choices" (Hewings and Hewings, 2002, p. 382; among others, the authors mention the adjectives *amazing*, *strange*, and *pointless*).

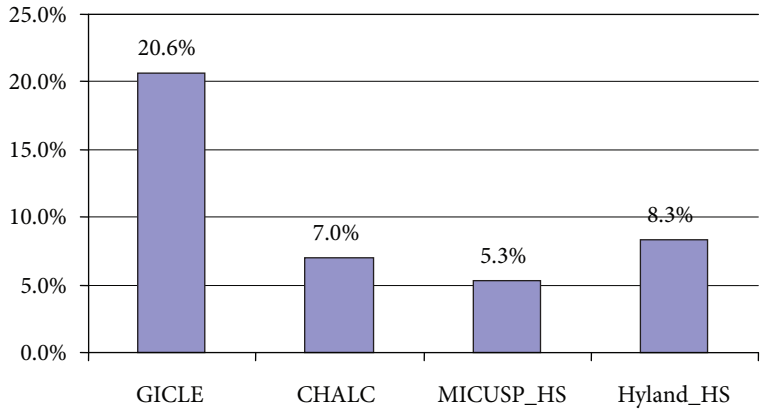


Figure 5. Proportions of adjective modification by boosters in introductory *it* patterns across corpora

To sum up the findings on introductory *it* patterns in apprentice and expert academic writing, I found that different patterns (*it is* (ADV) ADJ *to*-infinitive and *it is* (ADV) ADJ *that*-clause) correspond with different meanings and show different preferences in terms of adjective selection. This finding is very much in line with central Pattern Grammar statements on the relationship between patterns and meanings (see Section 2.2 above). I also found that learner/apprentice writers's uses of introductory *it* patterns deviate in many ways and to different extents from published/expert writing as captured in Hyland_HS. For instance, the proportions of the function most frequently expressed by the *it is* (ADV) ADJ *to*-infinitive pattern, 'difficulty', are higher in GICLE, CHALC, and MICUSP_HS than in Hyland_HS, whereas 'desirability' is less commonly expressed with this pattern by apprentice writers, in particular by those on more advanced levels, than by expert writers (see Figure 3 above). Interesting differences across corpora were also observed in the analysis of pattern-specific adjective selection. Novice academic writers seems to favour a small number of adjectives and use them considerably more often than Hyland_HS writers (e.g. *possible* and *important*) and at the same time avoid other adjectives which are frequent in Hyland_HS (*rational* and *interesting* in particular). The fact that some of the frequently used adjectives only

appear in the top-10 lists of two or three corpora (Hyland_HS and MICUSP_HS, MICUSP_HS and CHALC), may indicate that lexical preferences gradually shift as writers become more experienced. However, more evidence based on larger matching sets of data would be needed to strengthen this claim.

In Section 3.1 above, I described the corpora this study is based on and identified three variables that may have an influence on the results: (i) nativeness, (ii) general language proficiency, and (iii) expertise in academic writing. Since we do not observe a clear divide in the results between GICLE and CHALC (our non-native speaker corpora) on the one hand and MICUSP_HS and Hyland_HS (our predominantly native speaker corpora) on the other, nativeness does not appear to be the major factor that influences lexical-grammatical selection (at least when it comes to introductory *it* patterns; see also Wulff and Römer, forthcoming). Which of the two remaining variables is responsible for which of the identified differences between the datasets is at this point hard to determine, especially since general language proficiency and academic writing proficiency are difficult to tear apart. The fact that GICLE writer output differs in several ways from CHALC and from MICUSP_HS output may be accounted for by GICLE learners's lower general language proficiency but also by the smaller number of years they have spent at university. It should also be mentioned that, although in the selection of corpora a good attempt has been made to control for text type and discipline, differences in the results may partly be the consequence of not having been able to perfectly match disciplines and text types across corpora (e.g., CHALC contains essays and terms papers whereas Hyland_HS is made up of research articles). More research based on (to my knowledge yet unavailable) corpora that control for variables such as learner levels or years of academic instruction more systematically is badly needed. To be fully comparable, these corpora should ideally also cover the same text types in the same disciplines.

4. Conclusion

This paper has provided an overview of six selected research strands that focus on the lexis-grammar continuum and summarized some central insights gained by corpus linguists on how meanings are created in language. It has tried to show that core corpus linguistics centres around the interface of lexis and grammar and sees phraseology and phrasal units at the heart of language (to echo Ellis, 2008, p. 1; see also Sinclair, 2008, p. 407).

The paper has also tried to show that the research strands and theories presented here are not as diverse as they may seem at first, but have a number of things in common. Different camps in corpus linguistics have provided (and

continue to provide) massive evidence for the inseparability of lexis and grammar and for the close connection between patterns (or constructions) and meanings (or functions). As a case in point, the introductory *it* pattern and its subpatterns and lexical realisations in four corpora of apprentice and expert academic writing were discussed. I found clear associations between patterns, lexical items (in this case adjectives) and meanings, and a number of interesting deviations (in terms of frequencies, types, and functions of the patterns) for the examined apprentice writer corpora (GICLE, CHALC, and MICUSP_HS) from the expert norm (Hyland_HS). These deviations could not be explained on the basis of the nativeness vs. non-nativeness distinction (there is no clear divide between the two NNS corpora on the one hand and the two predominantly NS corpora on the other) but seem to be related to the two intertwined factors 'general language proficiency' and 'expertise in academic writing'. A thing that clearly emerged from the analysis is that learners and apprentice writers use patterns and within these patterns make lexical choices that are not in line with the target norm (successful/published writing) and sometimes even appear unusual or text-type inappropriate.

Based on the observations made in this paper, I would suggest that we respond to the observed deviations from the expert norm in general ELT and in EAP classes, both on introductory and advanced levels, and highlight for our students and novices in academia that lexical-grammatical patterns, collocations, lexical bundles, and constructions matter, thus helping them become accepted members of the specific community of practice they aim to belong to.

Notes

* I would like to thank Matthew Brook O'Donnell, Stefanie Wulff, and two anonymous reviewers for constructive comments on earlier versions of this paper.

1. The fact that these strands are dealt with under one heading is not supposed to suggest that they all have the same status in terms of being a theory (like Construction Grammar) or a methodology (like Collostructional Analysis) and are directly comparable. What connects the strands, is that they all take an integrated approach to lexis and grammar.

2. A related website at http://www.lexicalpriming.org (accessed 19 October 2008) provides information on the theory itself and links to related presentations by Hoey and his Liverpool colleagues.

3. This figure is given in Biber et al. (1999). Biber, Conrad and Cortes (2004, p. 376) take a more conservative approach and use a cut-off of 40 times per million words.

4. While not all Hyland_HS writers are native speakers of English, it can however be assumed that articles by non-native speakers have been checked and corrected by a native speaker. This is at least the policy of most of the journals from which the Hyland_HS articles have been taken.

5. Throughout the paper, frequencies are normed to counts per million words. Norming is a common procedure in corpus linguistics used to ensure that findings based on corpora of different size can be compared. While it would have been desirable to have matched corpora for this study in terms of size and number of texts, this was not feasible as such matched corpora of apprentice and expert academic writing are not currently available.

References

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

Biber, D. (2006). *University Language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

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

Biber, D., Conrad, S. & Cortes, V. (2004). *If you look at…*: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*, 371–405.

Cortes, V. (2004). Lexical bundles in published and student writing in history and biology. *English for Specific Purposes, 23*, 397–423.

De Cock, S. (1998). A recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English. *International Journal of Corpus Linguistics, 3*(1), 59–80.

Ellis, N.C. (2008). Phraseology: The periphery and the heart of language. In F. Meunier & S. Granger (Eds.), *Phraseology in Foreign Language Learning and Teaching* (pp. 1–13). Amsterdam: John Benjamins.

Fletcher, W. H. (2002–2007). *KfNgram*. Annapolis, MD: USNA. Available at http://kwicfinder.com/kfNgram/ (accessed 3 September 2008).

Francis, G., Hunston, S. & Manning, E. (1998). *Grammar Patterns 2: Nouns and adjectives*. London: HarperCollins.

Goldberg, A. (1995). *Constructions: A construction grammar approach to argument structure.* Chicago: The University of Chicago Press.

Goldberg, A. (2006). *Constructions at Work*. Oxford: Oxford University Press.

Goldberg, A. & R. Jackendoff. (2004). The English resultative as a family of constructions. *Language, 80*, 532–568.

Granger, S., Dagneaux, E. & Meuner, F. (Eds.). (2002). *International Corpus of Learner English. Handbook and CD-ROM.* Louvain-la-Neuve: Presses Universitaires de Louvain.

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

Gries, S.T. & A. Stefanowitsch. (2004). Extending collostructional analysis: a corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics, 9*(1), 97–129.

Groom, N. (2005). Pattern and meaning across genres and disciplines: An exploratory study. *Journal of English for Academic Purposes, 4*(3), 257–277.

Hewings, M. & A. Hewings. (2002). "It is interesting to note that…": A comparative study of anticipatory "it" in student and published writing. *English for Specific Purposes, 21*, 367–383.

Hilpert, M. (2008). *Germanic Future Constructions: A usage-based approach to language change*. Amsterdam: John Benjamins.

Hoey, M.P. (2004). Lexical priming and the properties of text. In A. Partington, J. Morley & L. Haarman (Eds.), *Corpora and Discourse* (pp. 385–412). Frankfurt: Peter Lang.
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. In U. Römer & R. Schulze (Eds.), *Exploring the Lexis-Grammar Interface* (pp. 33–47). Amsterdam: John Benjamins.
Hoey, M.P. & M.B. O'Donnell. (2008). Lexicography, grammar, and textual position. *International Journal of Lexicography, 21*(3), 293–309.
Hunston, S. (2008). Starting with the small words: Patterns, lexis and semantic sequences. *International Journal of Corpus Linguistics, 13*(3), 271–295. [Special issue on *Patterns, meaningful units and specialized discourses*, Eds. U. Römer & R. Schulze]
Hunston, S. & G. Francis. (2000). *Pattern Grammar. A corpus-driven approach to the lexical grammar of English*. Amsterdam: John Benjamins.
Hyland, K. (1998). *Hedging in Scientific Research Articles.* Amsterdam: John Benjamins.
Hyland, K. (2008a). Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics, 18*(1), 41–62.
Hyland, K. (2008b). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes, 27*, 4–21.
Leech, G. & S. Svartvik. (2002). *A Communicative Grammar of English (3rd edition)*. London: Longman.
Meunier, F. & S. Granger. (Eds.). (2008). *Phraseology in Foreign Language Learning and Teaching*. Amsterdam: John Benjamins.
Oakey, D. (2002). A corpus-based study of the formal and functional variation of a lexical phrase in different academic disciplines in English. In R. Reppen, S. Fitzmaurice & D. Biber (Eds.), *Using Corpora to Explore Linguistic Variation* (pp. 111–130). Amsterdam: John Benjamins.
Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.
Römer, U. (2007). Learner language and the norms in native corpora and EFL teaching materials: a case study of English conditionals. In S. Volk-Birke & J. Lippert (Eds.), *Anglistentag 2006 Halle. Proceedings* (pp. 355–363*)*. Trier: Wissenschaftlicher Verlag Trier.
Römer, U. & R. Schulze. (Eds.). (2008). *Patterns, Meaningful Units and Specialized Discourses*. (Special issue (13(3)) of the *International Journal of Corpus Linguistics*) Amsterdam: John Benjamins.
Römer, U. & R. Schulze. (Eds.). (2009). *Exploring the Lexis-Grammar Interface*. Amsterdam: John Benjamins.
Schmitt, N. (Ed.) (2004). *Formulaic Sequences. Acquisition, processing and use*. Amsterdam: John Benjamins.
Scott, M. & C. Tribble. (2006). *Textual Patterns. Key words and corpus analysis in language education*. Amsterdam: John Benjamins.
Sinclair, J. McH. (1987a). Collocation: A progress report. In R. Steele & T. Threadgold (Eds.), *Language Topics. Essays in Honour of Michael Halliday* (pp. 319–331*)*. Amsterdam: John Benjamins.
Sinclair, J. McH. (1987b). The nature of the evidence. In J.McH. Sinclair (Ed.), *Looking Up: An Account of the COBUILD Project in Lexical Computing* (pp. 150–159). London: HarperCollins.
Sinclair, J. McH. (1991). *Corpus Concordance Collocation*. Oxford: Oxford University Press.

Sinclair, J. McH. (1996). The search for units of meaning. *TEXTUS IX*, 75–106.
Sinclair, J. McH. (2004). *Trust the Text. Language, corpus and discourse*. London: Routledge.
Sinclair, J. McH. (2008). The phrase, the whole phrase and nothing but the phrase. In S. Granger & F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective* (pp. 407–410). Amsterdam: John Benjamins.
Sinclair, J. McH., Jones, S. & Daley, R. (1970/2004). *English Collocation Studies: The OSTI report*, R. Krishnamurthy (Ed.). London: Continuum.
Stefanowitsch, A. & S.T. Gries. (2003). Collostructions. Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics, 8*(2), 209–243.
Stubbs, M. (2009). Technology and phraseology: With notes on the history of corpus linguistics. In U. Römer & R. Schulze (Eds.), *Exploring the Lexis-Grammar Interface* (pp. 15–32). Amsterdam: John Benjamins.
Wulff, S. (2006). *Go*-V vs. *go-and*-V in English: A case of constructional synonymy? In S.T. Gries & A. Stefanowitsch (Eds.), *Corpora in Cognitive Linguistics. Corpus-based approaches to syntax and lexis* (pp. 101–125). Berlin: Mouton de Gruyter.
Wulff, S. & S.T. Gries. (this issue). *To-* vs. *ing*-complementation of advanced foreign language learners: Corpus- and psycholinguistic evidence.
Wulff, S. & U. Römer. (forthcoming). Becoming a proficient academic writer: Shifting lexical preferences in the use of the progressive. *Corpora.*

Author's address

Dr. Ute Römer
English Language Institute
University of Michigan
500 E. Washington St.
Ann Arbor, MI 48104
USA

uroemer@umich.edu

About the author

Ute Römer is currently Director of the Applied Corpus Linguistics Unit at the University of Michigan English Language Institute (http://elicorpora.info) where she manages the MICASE (Michigan Corpus of Academic Spoken English) and MICUSP (Michigan Corpus of Upper-level Student Papers) projects. Her primary research interests and areas in which she has published include corpus linguistics, phraseology, and the application of corpora in language learning and teaching. Her current research focus is on how corpus tools and methods can be used to identify meaningful units in specialized discourses.