**Proof corrections for**
Stefan Th. Gries: Corpus-based methods and SLA

- p. 411, par. par. 1 of Section 3: "to mean that the different parts of which the linguistic" → "to mean that the different parts of the linguistic"
- p. 412, last line before enumeration: "which are" → "which I take to be"
- p. 412, enumeration item 3: "refer to as argument structure constructions" → "refer to as (argument structure) constructions"
- p. 412, enumeration item 3: "(cf. Gries to appear b for" → "(cf. Gries to appear for"
- p. 413, par. 1 of Section 4: "or the clause/sentence" → "or the whole clause/sentence"
- p. 414: last par. before Section 4.1: "is dependent on how much data manual correction can be applied." → "is dependent on how much manual data correction can be done."
- p. 419: insert "(cf. also Wulff and Gries 2008 for a follow-up)" before the last period of Section 4.2.
- p. 421, penultimate par; "of various sorts and experimentation" → "of various sorts *and* experimentation"
- p. 422: middle of the page: delete the space between the word *output* and the period
- p. 423, par. 1: "Gries (to appear a)" → "Gries (2007)"
- p. 425, par. 2: "First, however, Baugh et al." → "First, Baugh et al."
- p. 427, note 10: "from the header in the" → "from the header (!) in the"
- p. 428: insert "Divjak, D. S. and St. Th. Gries. Ways of trying in Russian: clustering behavioral profiles. *Corpus Linguistics and Linguistic Theory* 2.1:23-60."
- p. 429: seven occurrences of "Gries, S. Th." → "Gries, St. Th."
- p. 429: "Gries, S. Th. to appear a. Exploring variability within and between corpora: some methodological considerations. *Corpora*." → ". Gries, St. Th. 2007. Exploring variability within and between corpora: some methodological considerations. *Corpora* 1.2:109-51."
- p. 429: "Gries, S. Th. to appear b. Phraseology and linguistic theory: a brief survey. In: Granger, S. and F. Meunier (eds.). *Phraseology: an interdisciplinary perspective*. Amsterdam and Philadelphia: John Benjamins." → "Gries, S. Th. to appear. Phraseology and linguistic theory: a brief survey. In: Granger, S. and F. Meunier (eds.). *Phraseology: an interdisciplinary perspective*. Amsterdam and Philadelphia: John Benjamins."
- p. 429: "Gries, S. Th. and D. Divjak. submitted. Behavioral profiles: a corpus-based approach to cognitive semantic analysis." → "Gries, St. Th. and D. Divjak. to appear. Behavioral profiles: a corpus-based approach to cognitive semantic analysis. In: Evans, V. and S. S. Pourcel (eds.). *New directions in cognitive linguistics*. Amsterdam and Philadelphia: John Benjamins."
- p. 430: "Stefanowitsch, A. and S. Th. Gries" → "Stefanowitsch, A. and S. Th. Gries"
- p. 431: insert "Wulff, St. and St. Th. Gries. 2008. *To-* vs. *ing*-complementation of advanced foreign language learners: corpus- and psycholinguistic evidence. Paper presented at the 15th World Congress of Applied Linguistics (AILA 2008), University of Duisburg-Essen; 24-29 August 2008."

**References missing from the reference section:**

Stewart, Dominic, Silvia Bernardini and Guy Aston 2004. Introduction. In: Aston, Guy, Silvia Bernardini, and Dominic Stewart (eds.). *Corpora and language learners.* Amsterdam and Philadelphia: John Benjamins, p. 1-18.
Van Hest, Erna. 1996. Self-repair in L1 and L2 production. Tilburg: Tilburg University Press.

**Another correction for the main text:**

On p. 418, line 3, the reference to Tono ("She") should be "He".
On p. 418, line 7-8, the reference to Tono ("her") should be "his".

# 16

# CORPUS-BASED METHODS IN ANALYSES OF SECOND LANGUAGE ACQUISTION DATA*

*Stefan Th. Gries*

## 1 Introduction

Second Language Acquisition (SLA) is a truly interdisciplinary field at the intersection of psychology, linguistics, applied linguistics, psycholinguistics, and educational science. Given the large number of associated fields and the fact that each of these fields has undergone considerable changes in a time period as short as the last 50 years, it is only natural that SLA is a diverse field in terms of both the theoretical approaches that have been adopted and the data and methodologies that have been applied. The present article looks at SLA at the intersection of the theoretical approach of Cognitive Linguistics—as is obvious from the title of this handbook—and the methodology of Corpus Linguistics—as is obvious from the title of the chapter. After a short introduction, I will first give a very brief overview of those characteristics of Cognitive Linguistics that are relevant in this connection. Then, I will discuss assumptions underlying contemporary Corpus Linguistics and at the same time highlight the large degree of overlap of the two fields. In the main section of this article, Section 4, I will discuss the main three methods of Corpus Linguistics and their application in SLA research. Section 5 will conclude and present some caveats and desiderata.

Let me begin by briefly clarifying a few things. First, some scholars distinguish between second language learning (L2 learning) and foreign language learning (FLL). In such cases, the former is used to refer to the learning of a language other than one's first language that takes place in a geographical/sociological context where the target language is spoken. The latter, by contrast, is used to refer to the learning of a language other than one's first language that takes place in a geographical/sociological

context where the target language is not spoken. In the present context, I will not distinguish the two settings because "the underlying learning processes are essentially the same for more local and more remote target languages, despite differing learning purposes and circumstances" (Mitchell and Myles 1998:1). Second, some scholars argue for a strict separation of SLA and language pedagogy, allowing for overlap only when "pedagogy affects the course of acquisition" (Gass and Selinker 2001:2). However, in the present context I will conflate these domains, though I will not be concerned with students' corpus explorations in second/foreign language learning contexts (e.g., data-driven learning).

Among the key questions that researchers in SLA attempt to answer are "how do learners create a new language system with only limited exposure (to the target language)?" and "what is learned, what is not learned, and why so?" Different schools of thought in SLA have proposed different answers to these questions. For example, behaviorist approaches viewed the acquisition of a language as the formation of new habits on the basis of development and reinforcement of stimulus-response pairs.

However, when behaviorist views of language came under attack, a more cognitive perspective was adopted both in psychology, where cognitive developmentalist views gained ground, and linguistics, where Chomsky's review of Skinner's *Verbal behavior* helped usher in a mentalist approach to psychology and linguistic theory by

1 demonstrating the rule-governed and creative nature of language acquisition;
2 pointing out the role of innate knowledge that aids children's acquisition process on the basis of impoverished input;
3 directing linguists' attention to putative linguistic universals, which also attest to genetically hard-wired linguistic knowledge.

For SLA, this implied a focus on the representation and acquisition of L2s rather than the way in which L2s are actually used (cf. Mitchell and Myles 1998:45), on whether or to what degree L2 learning processes have access to principles and parameters of Universal Grammar, and on transfer, e.g., whether or not parameter settings of L1 influence the acquisition of L2(s).

Most importantly for the purposes of the present volume, there are alternative approaches that could be broadly labeled as cognitively-oriented. These include (adopting the classification of Mitchell and Myles (1998)), but are not limited to, parallel distributed processing, functional models such as the Competition Model by Bates and MacWhinney (1982), and most relevant in the present context, Cognitive Linguistics, which will be discussed in the following section. As will become obvious

in Sections 2 and 3, many of the assumptions underlying cognitive-linguistic work—in particular the relevance assigned to frequency of patterns in learners' input—are not only gaining ground in SLA (cf., e.g., Ellis 2002a, b, 2006) but also are often completely analogous to working assumptions in Corpus Linguistics.

## 2 Cognitive Linguistics

Over the past 25 years, Cognitive Linguistics has matured into one of the most prominent alternatives to the linguistic paradigm of Chomskyan generative grammar. Cognitive Linguistics as such is no single theory but is probably best seen as a family of approaches which share several theoretical and methodological assumptions. The theoretical commitments lying at the heart of Cognitive Linguistics are the Generalization Commitment and the Cognitive Commitment (cf. Lakoff 1990). The former requires cognitive linguists to "characterize the general principles governing all aspects of human language" on all levels of description (e.g., phonology, syntax, semantics, pragmatics); this principle is basically a reformulation of a general scientific commitment to discover knowledge in the form of regularities. The latter Cognitive Commitment requires linguists "to make one's account of human language accord with what is generally known about the mind and the brain, from other disciplines as well as our own" (Lakoff 1990:40). The following exposition of the characteristics of Cognitive Linguistics will focus on the, to my mind, most fully articulated theory within Cognitive Linguistics: Langacker's Cognitive Grammar (cf. Langacker 1987, 1991). For reasons of space, I will restrict my discussion of the central properties of Cognitive Grammar to those that will be relevant for the discussion of corpus-based approaches in SLA.

In Cognitive Grammar, the only kind of element the linguistic system contains are symbolic units. A unit as such is defined as

> a structure that a speaker has mastered quite thoroughly, to the extent that he can employ it in largely automatic fashion, without having to focus his attention specifically on its individual parts for their arrangement [. . .] he has no need to reflect on how to put it together.
>
> (Langacker 1987:57)

A symbolic unit in turn is a unit that is a pairing of a form and a meaning/function, i.e., a conventionalized association of a phonological pole (i.e., a phonological structure) and a semantic/conceptual pole (i.e., a semantic/conceptual structure). There are essentially no restrictions on the nature and the number of formal elements that constitute the form,

and neither are there restrictions on the flexibility of the elements involved in the symbolic unit, but there are two major restrictions that need to be discussed. First, in order for something to attain unit status, the speaker whose linguistic system one is concerned with must have been able to form one or more generalizations (schemas in Langacker's parlance) which sanction the concrete instances. Crucially for our present purposes, the generalizations resulting in symbolic units can be made on the basis of any element from a continuum of increasingly abstract, or schematic, linguistic units. More specifically, generalizations can apply to and, thus, generate

1 maximally or highly specific elements such as morphemes or words;
2 intermediately specific elements such as partially lexically filled constructions; e.g., the *way*-construction (e.g., *He made his* way *through the crowd*), the *into*-causative (e.g., *She tricked him* into *marrying her*), or the conative construction (e.g., *He cut* at *the bread*);
3 highly abstract/schematic elements such as lexically unfilled constructions; e.g., the intransitive-motion construction (e.g., *The man was swimming in the ocean*), the ditransitive construction (e.g., *He gave her a book*), or linking constructions such as the subject-predicate construction.

Note in passing that this innocuous-sounding definition actually implies something very crucial, namely that Cognitive Grammar does away with a strict separation of syntax and lexis. This is because only symbolic units are allowed in the grammar, and generalizations across encounters of units can be made at the various levels of abstraction mentioned above. Thus, even syntactic patterns are meaningful in their own right, since they must have a semantic pole, which is highly schematic to allow for the diverse ways of how they can be instantiated. All this obviously stands in stark contrast to the various incarnations of transformational-generative grammar.

The second major restriction for something to attain the status of a symbolic unit is that the symbolic unit in question must have occurred frequently enough for it to be entrenched in a speaker/hearer's linguistic system.

These points have important corollaries. The first and most important one is concerned with the fact that Cognitive Linguistics in general and Cognitive Grammar in particular are usage-based approaches. This means that cognitive linguists assume that the use of linguistic elements and structures not only derives from the representation of the linguistic system in the minds of speakers but in turn also influences their representation via mechanisms of routinization, entrenchment, etc. More specifically, the frequency of use of a particular symbolic unit does not only bear on

whether some unit is represented in the system or not but also on how the unit is represented and how it is accessed and processed. There are two aspects to this, one concerned with token frequency, one with type frequency. As to the former, the assumption is that high token frequency correlates with strong entrenchment: the more often a speaker/hearer encounters a particular symbolic unit, the more entrenched this symbolic unit becomes in his linguistic system and, as indicated in the above quotation from Langacker, the more automatically the symbolic unit is accessed. As to the latter, the assumption is that the type frequency of a particular linguistic expression—the number of different tokens instantiating a particular symbolic unit—correlates with that particular symbolic unit's productivity. For example, the transitive construction can be argued to be more productive than the ditransitive construction because the former occurs with more and more varied verbs. Both of these aspects also reflect the fact that Cognitive Linguistics is a surface-oriented approach: rather than assuming that, for example, input to L1 acquisition is impoverished (in the sense of not providing the child with clues to the wide variety of empty categories and traces posited by generative linguists), cognitive linguists argue that that very same input does not contain all sorts of phonologically null elements but is rather rich in terms of probabilistic correlations or cues. This conception of language is particularly central to the input-driven models of learning and acquisition such as the Competition Model by Bates and MacWhinney (1982, 1989); cf. Ellis (2002a) for a comprehensive review. The acquisition of linguistic elements and structures, therefore, involves extensive frequency-based processing of actual linguistic input, processing that in turn involves pattern matching, bottom-up categorization and inferencing, and storing (of instances/exemplars and/or schemas, as will be described below).

The second major corollary is concerned with the degree of redundancy in the representational system. Since frequency of encounter is the only necessary condition for an expression's unit status (or sanctioning the unit status of a more abstract expression it instantiates), we have already seen above that speakers can make generalizations at many different levels of abstractness or, as Langacker calls it, schematicity. The point to be made here, however, is that these generalizations can be made on several levels *at the same time*. For example, encountering the word *years* will not only reinforce the entrenchment of the word *year* as a symbolic unit, the plural morpheme with [z] as its allomorph, and the schema that allows plural morphemes to be attached to nouns, it will, since *years* is among the most frequent plural nouns,[1] also reinforce the entrenchment of the word *years* as a symbolic unit. More broadly, while many frameworks would rule out such a redundancy in the storage system on grounds of lack of efficiency of storage, Langacker (1987:29, 42) argues forcefully against this so-called rule-list fallacy, pointing out that *years* can

be stored as a symbolic unit itself, too, if it is sufficiently frequent *even though* it could be derived productively from a regular plural formation rule. Crucially, this also applies to multi-word units: if a multi-word expression is encountered frequently enough—e.g., *I don't know*—it will be stored as a unit just like a monomorphemic word. To take a more abstract example, encountering the idiom *to spill the beans* will reinforce the entrenchment of all individual words and morphemes in the expression, the idiomatic unit (with its semantic pole "to reveal a secret"), and the transitive construction, which it also instantiates.

In a nutshell, just like many psycholinguistic models of language, Cognitive Linguistics presumes a variety of frequency effects on various levels. Put differently, there is a widespread recognition that absolute frequencies, relative frequencies (i.e., percentages), conditional probabilities, etc. are represented in the linguistic systems of speakers and, thus, play a primary role on all levels of linguistic analysis.[2] In addition, the last 25 years of both linguistics and psycholinguistics have seen an increasing reliance on the lexicon to investigate syntax (cf. Ellis 2002:157; Hudson, this volume). Given these two assumptions, it is therefore no surprise that Cognitive Linguistics is the theoretical framework that makes most use of the methodology of Corpus Linguistics. First, corpus linguists discarded a separation of syntax and lexis on independent grounds. Second, Corpus Linguistics is virtually exclusively based on frequency information—in fact, it is "the only reliable source of evidence for features such as frequency" (McEnery and Wilson 1996:12).

## 3 Corpus Linguistics

The expression *Corpus Linguistics* refers to a method in linguistics which involves the computerized retrieval, and subsequent analysis, of linguistic elements and structures from corpora. The concept of *corpus* in turn is a radial category with a prototypical central element. I will define a prototypical corpus as a machine-readable collection of (spoken or written) texts that were produced in a natural communicative setting, and the collection of these texts is compiled with the intention (i) to be representative and balanced with respect to a particular linguistic variety or register or genre and (ii) to be analyzed linguistically. In this definition, I use *representative* to mean that the different parts of which the linguistic variety one is interested in should all be manifested in the corpus. Relatedly, by *balanced* I mean that not only should all parts of which a variety consists be sampled into the corpus but also that the proportion with which a particular part is represented in a corpus reflects the proportion the part makes up in this variety and/or the importance of the part in this variety.[3]

From this characterization of Corpus Linguistics as involving "computerized retrieval," it follows that, strictly speaking at least, the only

thing corpora can provide is information on frequency, be it frequency of occurrence (of morphemes, words, grammatical patterns, . . .) or frequencies of co-occurrence of the same kinds of items (as determined by the computerized retrieval). The assumption underlying basically all corpus-based analyses, however, is that formal differences correspond to functional differences such that different frequencies of (co-)occurrences of formal elements are supposed to reflect functional regularities, where functional is meant here in the broadest sense including both semantic and pragmatic aspects of linguistic behavior.

This fact alone—that Corpus Linguistics is basically all about frequencies—would already provide for a strong affinity of Cognitive Linguistics and Corpus Linguistics: Corpus Linguistics provides exactly the kind of data that are at the heart of Cognitive Linguistics. However, the overlap of many of the two approaches' core assumptions and notions is much larger, in fact so large that it is surprising that this is not usually recognized more explicitly (cf. Schönefeld 1999 for a laudable exception). For example, we have seen that the only element in Cognitive Grammar is the symbolic unit, which can take on differently schematic forms. These are in fact perfectly reflected by the main ontological elements in Corpus Linguistics, which are

1 words, which are symbolic units;
2 phraseologisms, which are often defined as co-occurrences of at least one word form and other elements (i.e., collocations or colligations) which function as a semantic unit and, thus, often correspond to the partially lexically-filled constructions from above;
3 syntactic patterns, which correspond to lexically unfilled, highly schematic symbolic units, which Construction Grammarians in Cognitive Linguistics refer to as argument structure constructions (cf. Gries to appear b for a more detailed discussion of these parallelisms).

Just like Cognitive Grammar, corpus linguists also use sufficient frequency of occurrence as the usual definition for something to count as a phraseologism (Hunston 2002: Chapter 6) or a pattern (Hunston and Francis 1999:37). In line with this inventory of elements, corpus linguists have also abandoned the division of syntax and lexis. What is more, we have seen above that the notion of a (symbolic) unit in Cognitive Grammar entails that units are accessed automatically, i.e., fast and without the need to analyze their internal structure. Exactly these notions figure in the formulation of one of the most prominent principles in contemporary Corpus Linguistics, Sinclair's so-called idiom principle. This principle states that "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 analyzable into segments" (Sinclair

1991:110) and contrasts with the open-choice principle, which states that "[a]t each point where a unit is completed (a word or a phrase or a clause), a large number of choices opens up and the only restraint is grammaticalness" (Sinclair 1991:109). In other words, corpus-linguists share Langacker's rejection of the rule-list fallacy and claim—cf., for example, Pawley and Syder (1983:213, 215ff.)—that speakers' mental lexicons do contain much more than just lexical primitives, namely also probably hundreds of thousands of prefabricated items that could be productively assembled but are, as a result of frequent encounter, redundantly stored and accessed.

## 4 Three corpus-linguistic methods in SLA

For the purposes of the present review, it is useful to distinguish three different kinds of corpus-linguistic methods. First, there are frequency lists and collocate lists, or collocations. These constitute the most decontextualized methods, with the possible exception of one search expression largely ignoring the context in which an utterance or a sentence was produced. Second, there are colligations and collostructions, in which context is largely reduced to co-occurrences of lexical elements with a particular grammatical element or structure. Finally, there are concordances (of search expressions), which usually provide the occurrence of a match of the search expression in a user-defined context window, often four words to both the left and the right of the match or the clause/sentence in which the match occurred.

These methods will be discussed in more detail in the remainder of this section. They can be applied to several different kinds of data of interest to the SLA researcher. Depending on what kinds of corpora are available, one can look at all the kinds of language that will influence the output of the learner:

1 how does the input language pattern?
2 how does the native language of the learners pattern?[4]
3 how does the target language pattern?
4 what are the differences between how the native language and the target language pattern? (That is, all sorts of comparisons between these, cf. contrastive analysis.)

On the other hand, one can directly look at the output of the learner:

1 how does the interlanguage pattern? and/or
2 which (kinds of) errors do the language learners commit (computer-aided error analysis)?

Obviously, in many studies information from various of the above sources will, and actually should, be combined in order to be able to isolate sources of observed effects (cf. Juffs (2001:312) for a similar argument and, e.g., Tono's (2004) or Borin and Prütz (2004) for applications).

One final word of caution. Corpus-linguistic methods can basically be applied to symbolic units of all kinds of schematicity, but given the current state of the art in corpus annotation, not all symbolic units are equally amenable to corpus-based analysis. More specifically, symbolic units differ with regard to the precision and the recall with which they can be retrieved for analysis automatically. Symbolic units of low degrees of schematicity—most morphemes and words—are often easy to retrieve automatically and do not even require corpora with specific annotation (cf. Section 5 for some caveats, though). In other words, their retrieval is characterized by high precision and high recall. Symbolic units of intermediate schematicity are more problematic. In the absence of well-annotated corpora, one usually must conduct lexical searches and then weed out false hits manually. This way, the retrieval achieves high precision, but the recall is dependent on how much data manual correction can be applied. If the amount of lexically-based hits makes an exhaustive correction impossible, recall will be low—alternatively, recall can be high, but the effort required may increase dramatically, and I know of cases where many thousands of matches were combed through manually. Symbolic units of high schematicity such as argument structure constructions, finally, often require part-of-speech (POS) tagged if not syntactically annotated corpora for automated retrieval. If these are not available, some schematic symbolic units, those that contain at least one lexical element, can be retrieved with high recall, but require extensive manual post-editing to yield acceptable degrees of precision.

### *4.1 Frequency lists and collocates / lexical co-occurrences*

The most basic corpus-linguistic tool is the frequency list. In the most basic sense, a frequency list indicates how frequent words are in a corpus. Usually, a frequency list is a two-column table with all words occurring in the corpus in one column and the frequency with which they occur in the corpus in the other column. Typically, one out of three different sorting styles is used: alphabetical (ascending or descending), frequency order (ascending or, more typically, descending), and first occurrence (each word occurs in a position reflecting its first occurrence in the corpus). While this is certainly the most widespread kind of frequency list, theoretically other frequency lists are conceivable; depending on the corpus makeup and annotation, one can find frequency lists of morphemes, reverse frequency lists of words (to, say, group together all regular adverbs ending in *-ly* in English), pairs or even larger uninterrupted chains of words,

interrupted sequences of words, lemmas (as opposed to word forms), parts of speech, syntactic patterns, constructions, etc.

Frequency lists of various kinds of units or constructions are useful for a variety of purposes in the domain of SLA. In a pedagogy context, frequency lists are mostly built on the assumption that it is more useful for L2 learners to learn first those units that are particularly important in the target register/genre/variety, and not surprisingly in Corpus Linguistics, "particularly important" translates into "particularly frequent." This assumption can be found in virtually all contemporary introductions to Corpus Linguistics and has, for example, been argued for by Sinclair and Renouf (1988:148): "the main focus of study should be on (a) the commonest word forms in the language; (b) the central patterns of usage; (c) the combinations which they usually form"; cf. also Biber and Reppen (2002). However, frequency lists are also theoretically relevant in Cognitive Linguistics because, as outlined above, the more frequent a linguistic expression, the more entrenched it is assumed to be and the more likely it has unit status.

On the basis of the above assumption, the most straightforward application of frequency lists in SLA has been in the area of syllabus or curriculum development, where frequency lists are used to determine how much it is that is to be learned (cf. Hazenberg and Hulstijn 1996), what to attend to first and foremost, and in what order, ideally focusing (first) on what is typical/atypical rather than on what is possible/impossible. The logic is that since frequencies of symbolic units typically exhibit a Zipfian distribution—very few types account for very many tokens—the benefit of learning the most frequent elements first should be enormous (cf. Willis 1990 for some statistics and detailed discussion).[5] For example, Grabowski and Mindt (1995) provide a frequency list of irregular verbs that, if used in L2 teaching, would allow learners to quickly account for more than 80% of the verb tokens in the corpora analyzed. Similar proposals have been made for a variety of phenomena: modal verbs (Mindt 1995), markers of epistemic modality (Holmes 1988), progressive aspect (Römer 2005), etc. A more refined strategy—more refined in the sense that this strategy is statistically more sophisticated and allows for register/genre specific lists—is the use of key words, where key words are defined as those words in a corpus which are (significantly) overrepresented in this corpus as revealed by a statistical comparison to a (usually much larger and more overall representative) reference corpus (usually by means of chi-square of log-likelihood statistics).

Another relevant aspect in the more teaching-oriented area of the discipline is that of using frequency lists to quantify and/or compare the attainment of language proficiency. The most basic and most frequent statistic employed in this context is certainly the type-token ratio. For example, Cadierno (2004) analyzes how native and non-native speakers

express motion events differently in verb-framed and satellite-framed languages on the basis of type-token statistics of motion verbs in L1 and L2 Spanish. However, the problematic nature of this statistic is by now well known (cf. Granger and Wynne 1999) and other, more sophisticated indices of lexical richness such as the Lexical Frequency Profile (cf. Laufer and Nation 1995, and Meara 2005 for a critique) have been developed to serve as indicators of the development of vocabulary (cf. also Read 1997, 2000; Baayen 2001; Chipere, Malvern, and Richards 2004).

As to more theoretical approaches, frequency lists are used as approximations of the frequencies of elements in the input to the L2 learner or in the target language—"approximations" because it is of course unclear to what degree corpus frequencies can in fact reflect actual learner input frequencies (cf. Ellis and Schmidt 1997), and the same applies of course to the reflection of target-language frequencies. Also, frequency lists serve to pick experimental stimuli: for example, Wolter (2001) uses experimental stimuli chosen on the basis of a word-frequency list generated from the Collins Cobuild Bank of English.

While most of the above work focuses on individual lexical words and syntactic patterns, there is also some interesting work involving frequency lists of elements consisting of more than just one word or one syntactic pattern, namely co-occurrences of words, variously referred to as collocations, lexical *n*-grams (where *n* refers to the number of words involved), multi-word units, prefabs, and certainly other expressions. For example, De Cock et al. (1998, also De Cock 1998) compare "prefabs" in native-speaker and learner corpora to test the hypothesis that learners tend not to use formulae as frequently as native speakers do. Biber et al (1999:993–994) examine "lexical bundles" in conversation and academic prose. Howarth (1998) reports quantitative results regarding the collocational density and stylistic conventionality of non-native-speaker collocations and idioms as well as qualitative results regarding the errors made, and strategies used, by non-native speakers. Other work is somewhat more abstract. For example, Aarts and Granger (1998) do not focus on *n*-grams of words but rather compare frequencies of *3*-grams of part-of-speech (POS) tags in the ICLE interlanguage corpus with the corresponding frequencies in a corpus of L1 speakers of the target language English, an instance of contrastive interlanguage analysis. Borin and Prütz (2004) go even further and also include *n*-gram frequencies of POS tags in the native language of the language learners.

Finally, it is worth noting in passing that such raw frequency statistics are often reported at the beginning of results sections to give a feel for the kinds of corpora that were used or the kinds of elicited corpus results that were obtained (cf. Waara's (2004) analysis of the argument structures taken by *get* for an example).

### *4.2 Colligations and collostructions: lexico-grammatical co-occurrence*

The next method includes more contextual information than just frequencies or collocations and bridges the gap to grammatical analysis. Much of the work in Cognitive Linguistics—especially work in Construction Grammar—is concerned with colligations and collostructions, the co-occurrence of lexical and grammatical elements.[6] More specifically, *strong tea*—as opposed to *powerful tea*, to use a famous example—is a collocation as is *hermetically sealed* since the two co-occurring elements, *strong* and *tea* as well as *hermetically* and *sealed*, are all lexical items. By contrast, the fact that the verb *to hem* is usually used in the passive is a colligation or collostruction since only one of the co-occurring elements is a lexical item whereas the other is a grammatical element, the passive construction.[7] Recall, however, that, since Cognitive Grammar does away with a strict separation of syntax and lexis, collocations as well as colligations and collostructions are all co-occurrences of symbolic units, if only at different levels of schematicity, which as pointed out above, usually goes hand in hand with an increase of the time and effort required for retrieval and annotation. Two perspectives are conceivable: the analysis starts out either from the retrieval of the lexical element of the colligation/collostruction or from the grammatical element. In what follows, examples of both strategies will be discussed.

One of the most important areas of application for colligations and collostructions would obviously be the investigation of grammatical proficiency as such, as well as the knowledge of interdependencies of lexis and syntactic patterns / constructions. The most central domain of study in this area is probably verb subcategorization patterns. A huge research project concerned with charting out the ways in which syntax and lexis interact in the usual target language, English, is Hunston and Francis's pattern grammar project based on the COBUILD project. Interestingly, many of their conclusions could in fact be straight quotations from studies independently arrived at in Construction Grammar, as the following quote exemplifies:

> [. . .] words sharing the same patterns tend to fall into groups based on shared aspects. This in turn suggests that *the patterns themselves can be said to have meanings*, and there is some evidence that the use of a lexical item with a pattern it does not commonly have is a resource for language creativity and, possibly, for language change.
>
> (Hunston and Francis 1998:69; my emphasis, STG)

A study more directly concerned with acquisition patterns is the



above-mentioned paper by Tono (2004), who conducts a multifactorial analysis of the factors governing the acquisition of verb subcategorization frame patterns by Japanese learners of English. She uses three corpora: an interlanguage corpus of Japanese L2 learners of English, an L1 speaker corpus of Japanese, and a target language corpus, which is a corpus of textbook English rather than "real, authentic" English, an unfortunate choice in her case, at least from my point of view (but cf. Tono (2004:51f.) for a defense of her choice of a textbook corpus as representing the target language).

An example closer to a Cognitive Linguistics / Construction Grammar approach is Waara (2004). She attempts to compare argument structure constructions of the English light verb lemma *to get* on the basis of two corpora of test interviews—one of L1 speakers of English, one of Norwegian L2 learners of English. She starts out from the forms of the verb and then reports the frequencies with which different argument structure constructions occur with *to get* aiming to explain how learner constructions reflect different transfer, blending, and overgeneralization of the developing L2 system.

Colligations and collostructions also provide approximations to input frequencies and target language frequencies of lexico-grammatical co-occurrences in much the same way as frequency lists and collocations do for words. For example, several central notions of the Competition Model—cue validity, cue availability, and cue reliability—are usually approximated probabilistically on the basis of co-occurrence frequencies of some cue(s) (such as word order, agency, animacy, . . .) and some grammatical structure or some comprehension preference under consideration (cf., e.g., Kempe and MacWhinney 1998); these notions are very similar especially to Hoey's more recent usage of *colligation* (cf. again Note 7).

An interesting approach—from my certainly not unbiased point of view—to verb-subcategorization preferences and constructions of L2 learners is the study by Gries and Wulff (2005). They investigate whether advanced German learners of English have acquired knowledge of the syntax-lexis interface that is similar to that of native speakers of English. They use data from an L1 English corpus (the British component of the International Corpus of English) and verb-subcategorization preferences obtained from a parsed L1 German corpus and correlate these corpus data with results from a syntactic-priming and a sorting experiment conducted with the German learners. Syntactic priming refers to the tendency to re-use syntactic patterns used shortly before. In the syntactic priming experiment, subjects were asked to complete sentence beginnings. Some sentence beginnings served as primes for ditransitive constructions (such as *The racing driver showed the helpful mechanic . . .*) while others served as primes for the prepositional dative (such as *The racing driver showed the*





*torn overall . . .*). The dependent variable was the choice of syntactic construction upon presentation of a non-biasing sentence fragment after the prime (such as *The angry student gave . . .*). In the sorting experiment, the German learners were given 16 cards, each displaying one sentence, and asked to sort these into four groups of four sentences each. The sentences crossed four verbs (*cut*, *get*, *take*, and *throw*) and four argument structure constructions (the transitive, caused-motion, resultative, and ditransitive construction). The dependent variable was whether the subjects adopted a perceptually simpler verb-based sorting style or a more complex construction-based sorting style. Their findings lend support to a conception of L2 learning that involves the development of a probabilistic system from (probabilistic correlations in) actual input and is, thus, fully compatible with a usage-based approach.

First, they find that advanced German L2 learners do not only exhibit overall syntactic priming effects in English comparable to those of native speakers of English but they also exhibit verb-specific priming effects that are (i) very highly correlated with the verbs' subcategorization preferences in native speaker English and (ii) completely unlike the verbs' German translation equivalents' subcategorization preferences in native speaker German. Second, they find that (different) advanced German L2 learners also exhibit a preference for sorting sentences according to the argument structure constructions they instantiate rather than according to the verb they feature. Interestingly, the grouping of constructions conform to a Construction Grammar account of how these constructions are related to each other. What is more, one can now compare the average number of cards one has to move to arrive at purely construction-based sortings (where smaller values indicate a higher degree of construction-based sorting) from several studies: Bencini and Goldberg (2000:645) obtained a value of 3.2 for native speakers; Gries and Wulff's (2005:192) replication with very advanced German learners of English obtained a value of 3.45; Liang's (2002) replication with intermediate Chinese learners of English obtained a value of 4.9. Obviously, the degree of proficiency of the L2 learners of English is well reflected, at least on an ordinal scale, by their recognition of, and sorting based on, argument structure constructions in English; a finding which is not only interesting for the above theoretical reason but also because it at the same time demonstrates the power of converging evidence from different experimental studies and corpus-linguistic data.

### *4.3 Concordances*

Concordances are the most comprehensive tool in the sense of providing the most informative context for the matches of the search expression(s). Usually, a user specifies a search expression, which can, depending on the

annotation of the corpus, involve information from many different levels of linguistic analysis, e.g., lemmas, POS tags, etc. In addition, the user defines the format of the output by typically either specifying a number of words around the match that should be displayed (usually referred to as window or span) or by setting the concordancer to display the whole sentence in which the match occurred (or even additional sentences). This way, concordances do not miss out on any information and are probably the most universally applicable and widespread method; only a tiny snapshot of what is possible beyond the above examples can be given here.

Cadierno (2004) applies Slobin's notion of thinking-for-speaking as well as Talmy's typology of motion verbs and investigates how native speakers of a satellite-framed languages describe events in a verb-framed language, comparing data from narratives elicited from L1 speakers of Spanish and Danish L2 learners of Spanish. She finds how L2 proficiency and cross-linguistic influence interact in the expression of motion events: on the one hand, Danish learners of Spanish "exhibited a relatively higher degree of complexity and elaboration of the semantic component of path of motion," but on the other hand, they "did not transfer the characteristic typical typological pattern of the L1 into the L2" Cadierno (2004:41ff.).

Nesselhauf (2004) investigates the errors committed by German learners of English in the use of support-verb constructions with *make*, *have*, *take*, and *give*. Given that her corpus—a part of the German section of the ICLE—is not annotated syntactically, her retrieval of support-verb constructions was a two-step procedure such that she first retrieved all matches of the verb forms and then manually identified verb-support constructions. Her conclusions are interesting in that she cautions against teaching recommendations exclusively based on the criterion of frequency. A study with a slightly similar focus by Altenberg and Granger (2001) looks at the use of the high-frequency verb *to make*, comparing data from interlanguage and native corpora.

A not yet particularly widespread issue in SLA is that of contrastive semantic prosody. Semantic prosody refers to aspects of typically evaluative meaning words take on from their most frequent collocates; for example *happen* and *set in* usually take collocates referring to unpleasant situations. Accordingly, contrastive semantic prosody refers to the different semantic prosodies of translational equivalents in different languages. For example, Xiao and McEnery (2006) explore concordance data for collocations that reveal semantic prosodies in the target language and the interlanguage to determine the degree to which language learners master semantic prosodies; they also point out the importance of making teachers and learners aware of cross-linguistic and text type-specific differences.

Many other approaches are conceivable since concordances basically

impose no limit on the amount of information that can be included. Thus, even issues that superficially seem to be less tied to formally identifiable patterns, for example the polysemy of prepositions, metaphorical uses vs. literal uses of verbs such as *to see*, quantification in L2 (cf. Kennedy 1987), etc.—all of these aspects await more attention from corpus linguists.

## 5 Caveats, conclusions, and desiderata

This handbook chapter has discussed the intimate relation of Cognitive Linguistics and Corpus Linguistics as well as how many examples of cognitive-linguistic concepts and corpus-based methods are used in SLA research.

While the above discussion has—for expository reasons—treated different methods separately, the theoretical ideal would of course be that, for example, detailed studies of the behavior of any symbolic unit integrate information from all these methods. Hunston (2002: 76ff.) gives an excellent example in her textbook discussion of the word *leak*. She first finds that collocates of the verb *to leak* indicate that the verb has at least two broader senses. One is concerned with the movement of a liquid through objects that are supposed to be solid and reflected by collocates such as *oil*, *water*, *gas*, and *roof*; the other, more metaphoric, sense is concerned with making information available as reflected by collocates such as *document*(*s*), *information*, *report*, *memo*, and *confidential*. However, she then also shows that the collocation frequencies alone cannot reveal that the metaphoric sense is used both intransitively and transitively, something that can only be noticed by looking at syntactic patterns or the concordance lines.

Also, it is clear that the larger the number of different corpora that are included in one's study, the greater the likelihood of robust and revealing findings: as the work by, say, Tono (2004) shows, including as many as possible of the different kinds of corpora discussed at the beginning of Section 4 into a multiple comparison approach will assist in separating L1-induced variation from L2-induced variation from learner input-induced variation. Finally, it is all too obvious that probably every kind of SLA research can benefit from converging evidence from several methods such as corpus data of various sorts and experimentation, and a few such examples were mentioned above.

An advantage of corpus data seems to be that once the required corpora are available, frequency lists, concordances, collocates, etc. seem easy to come by. For example, there are many corpus programs available with which frequency lists of corpora are just a few mouse clicks away;[8] the situation is similar with regard to collocate displays (cf. Wiechmann and Fuhs 2006 for a comparative review). However, there are also some problems that are associated with these methods that tend to be underestimated

and not addressed explicitly enough in much recent work. One is concerned with the fact that most of the data retrieval is done by computer programs. For example, frequency lists, collocations, and concordances presuppose that the retrieval is done correctly and that one has a definition of what a word is (and, for the sake of comparison with other work, that this definition is shared by other linguists and their computer programs).[9] However, this may not be so. Concordance programs usually define word forms as a sequence of alphabetic (or alphanumeric) characters uninterrupted by whitespace (i.e., spaces, tabs, and newlines), but some also allow the user to specify what to do with hyphens, apostrophes, and other special characters. For example, concordancers may differ as to how they handle "better-suited" or "ill-defined"; "armchair-linguist," "armchair linguist," and "armchair-linguist's"; "This" and "this"; "1960"; "25-year-old"; and "favor" and "favour." Thus, one needs to exercise extreme caution when comparing frequency lists from different sources. In addition, some corpus files come with a lot of annotation in headers and not every program is equally good at ignoring the words in the header etc.[10]

More often than not, it is the restricted accessibility of SLA corpus data that puts considerable limits to the kinds of studies that are possible or not; however, with regard to those corpus data that actually are widely accessible to date, the very fact that they are so easily accessible may also be problematic. For example, a potential downside of concordances derives from their strongest advantage, namely the fact that they provide the most informative output . This high degree of informativeness can be disadvantageous because it usually goes in hand in hand with not being able to automate (parts of) the analysis. For example, if all one is interested in is the number of different prepositions used after a verb in the same clause, it is easy to retrieve this information largely automatically with a snippet of code of a programming language. However, if one really wishes to exploit all the information available in concordances of, say, a particular verb, then there are many characteristics that computer scripts usually fail to retrieve with a high degree of precision; examples include animacy of participants, clause types, transitivity of the verb, properties of the process denoted by the verb, the metaphoricity of the use of the verb, etc. The number of properties that need to be coded manually or at least semi-manually can easily reach 100,000 or more (cf. Gries 2006 or Divjak and Gries 2006). Thus, often the data to be included in an analysis are reduced such that only every *n*-th match of an actual concordance output can be investigated in as much detail as would usually be desired. Therefore, adequate sampling strategies are of vital importance in this context. One possibility would be to use stratified sampling on the basis of top-down distinctions, sampling, for example, different speaker groups (e.g., as defined by perceived proficiency), different speakers, different genres, etc. Another possibility, which has so far not been used widely in

Corpus Linguistics, is to apply resampling and/or bootstrapping techniques to better (i) derive the relevant sample groups directly from the data themselves, i.e., in a bottom-up fashion, without necessarily invoking any particular researcher's preconceptions, and (ii) quantify the variability in the data more precisely than simple summary statistics do; cf. Meara (2005) and Gries (to appear a) for discussion and exemplification.

Yet another issue is concerned with the kind of frequencies that are provided. There are three aspects to this. First, much work using frequencies exhibits a—from my point of view—rather unfortunate preference to use only raw frequencies of occurrence or of co-occurrence in a usually arbitrarily defined span or of a usually arbitrarily defined length. Second and relatedly, the frequencies that are obtained this way are often just reported or evaluated in but the simplest possible ways. Third, it seems as if there is as yet no rigorous operationalization of when something is frequent enough to be considered a unit in the above sense of the term. Rather, different scholars just use subjective impressions of what is frequent enough or not, and, to quote but one example, Hunston (2002:147) even states pessimistically

> How many examples of a three-, four- or five-word sequence are necessary for it to be considered a phrase [sic! I guess what is meant is "a phraseologism"; STG]? As this is not an answerable question [. . .]

However, there is a lot of research that is just waiting to be exploited in Cognitive Linguistics / Corpus Linguistics in general and in corpus-based SLA in particular. As to the first, there is immensely useful work on, for example, how to determine the ideal span size or slot to investigate in terms of collocations. For example, much work involving collocations is based on the words found in a span of three to five words around the node word that was retrieved. However, this leaves much noise in the data to be filtered out. An interesting refinement is, for example, Mason's (1997, 1999) work on what he refers to as lexical gravity, namely a method to identify those positions around a node word which exhibit the smallest amount of entropy and are, thus, most revealing for the subsequent analysis. Another example is work by Kita et al. (1994) and many others, who have provided interesting proposals concerning the identification of multi-word units in corpora.

As to the second and related aspect of evaluation of frequencies, there is probably even more previous work on how statistical approaches going beyond absolute or relative frequencies allow for a more appropriate analysis of the data. For example, Gries, Hampe, and Schönefeld's (2005, to appear) work on the so-called *as*-predicative shows that collocational statistics (such as the Fisher-Yates exact test) outperform raw frequencies

in terms of their predictive power in experimental setups. For example, work by Evert and his collaborators (cf. Evert and Krenn 2001) investigates which measures of collocational attraction are most useful in terms of reflecting native speaker intuitions, etc. Also, there are quite a few, but not yet enough, studies that exhibit the degree of statistical sophistication that a discipline whose main subject matter is frequencies would lead one to expect; for example, Tono (2004) or Borin and Pütz (2004) are laudable exceptions in this regard, using multifactorial and other significance tests.

As to the third aspect regarding frequency thresholds, I am not in a position to propose a universally applicable frequency criterion for unit status—perhaps some ratio of observed vs. expected frequency of occurrence will be useful—but there is work approaching this issue in interesting ways. For example, studies such as Bybee and Scheibman (1999) or Jurafsky et al. (2001), in which the assumption that particular expressions with high-token frequencies have unit status is supported by additional independent evidence concerning an expression's status as an autonomous unit, namely the readiness of these expressions to undergo, say, processes of grammaticalization and/or phonological reduction (e.g., the reduction of the vowel in *don't* in *I don't* or *why don't you . . .?*). While these are promising steps, this is certainly still a crucial issue that needs to be addressed and developed further if the notion of unit status is to be more than a guesstimate. In addition, while the quoted studies demonstrate the relevance of frequency data at the level of the word forms (e.g., *don't* or *that you*), there may well be occasions on which other levels of granularity—e.g., the lemma—are even more revealing (cf. Harrington and Dennis 2002 on what they call the redescription problem). For example, inspecting spoken American corpus data shows that *you know what I mean?* may be a unit, as may *you know what I'm sayin'?* or *you know what VP?* Especially from an exemplar-based theoretical perspective, speakers/learners may make generalizations with different degrees of predictive power on many levels at the same time. It remains to be hoped that these and other methodological advances will further our use and understanding of frequency data.

In the area of syllabus design, it is not always clear what corpora to choose as reference corpora for the target language (cf. again Tono 2004:51ff. for discussion) and how useful frequency lists that do not take register differences or communicative goals of speakers/writer into account can actually be. Perhaps even more fundamental, however, is the question of if and to what degree frequency-based syllabi are more useful to L2 learners in the first place. While it may seem intuitively obvious that (i) frequent units are more useful to learners than rare units and that (ii) teaching material should resemble authentic speech, there are also several not-so-obvious issues involved that are not always topicalized. First,

one has to decide on which level of granularity frequency is supposed to be important. On the one hand, one could go for the most frequent lexical forms (or, of course, syntactic constructions, etc.). On the other hand, one could go for the most frequent lexical lemmas. Yet again, one could go for the most frequent senses of lexical lemmas or for senses in particular syntactic patters, etc. Different level choices may yield different results and it is not immediately obvious which level of granularity is most useful.

Second, even if one decides on a particular level of granularity, it is still not always clear whether the learner benefits more from the exposure to authentic examples. One may argue that the exposure to authentic examples increases the likelihood of the learner developing a network of differently weighted probabilistic connections that approximates that of a native speaker. However, proponents of corpus-based material face several challenges. First, however, Baugh et al. (1996:43) argue that "[m]ost citations are unsuitable for a learner dictionary because they are too complex grammatically, contain unnecessary difficult words or idioms, or make culture-dependent allusions or references to specific contexts," and many reference works have therefore chosen to carefully edit authentic examples. Second, authenticity does not automatically entail typicality, and it may well be the case that learners benefit equally or more from the careful comparison of, say, minimal-pair-like examples that have been constructed to highlight a particular aspect or contrast to be learned. In other words, the issue is whether the saliency created in constructed examples may outperform the frequency of authentic examples. It is worth pointing out in this connection that authors such as Nesselhauf (2004), Mauranen (2004), and Gabrielatos (2005) underscore the importance of including not just frequency-based factors into attempts at improving teaching material. On these grounds, and because it is my impression that so far even some of the most ardent defenders of the authenticity/frequency camp—e.g., Glisan and Drescher (1993) or Römer (2004)—have not yet provided experimental evidence to bolster their claims as to the utility or even indispensability of corpus-based curricula, the issue of whether corpus frequency-based materials are necessary or useful is still largely unresolved.

In spite of all these caveats and desiderata, it should be obvious that corpus-based work has a lot to offer to the analyst. I therefore hope and predict that more and more SLA corpora will be developed, annotated, and shared among research groups—especially corpora on interlanguages other than English—and that methods to increase precision and recall of symbolic units of all kinds will be developed to help cognitive- and corpus-linguistics approaches to SLA mature and prosper.

## Notes

* I am very grateful to Stefanie Wulff for a lot of advice and discussion. Also, I thank the editors of this handbook for their detailed comments and Beate Hampe for feedback. The usual disclaimers apply.

1 Note, for example, that it is the most frequent word tagged as a plural noun in the British National Corpus World edition.

2 Cf. Bley-Vroman (2002) or Eubank and Gregg (2002) for commentaries that accord frequency a less prominent role for data analysis and theory development.

3 Note how his definition already implies that usually SLA corpora are not prototypical corpora—according to the above definition, that is. This is not meant to imply they are deficient but that they are specialized in one ore more respects. For example, a widely known SLA corpus, the International Corpus of Learner English (Granger, Dagneaux, and Meunier 2002) largely consists of student essays and it is debatable, to say the least, whether essays students write at the request of a language instructor instantiate communication in a natural setting (cf., e.g., Granger 2002:7f. and Stewart et al. 2004: Section 4 for discussion); similar remarks apply to the Longman Learners Corpus (1993). Often, SLA research involves experiments which could be viewed as corpus-generation experiments. For example, van Hest (1996) uses a corpus of self-repairs in Dutch in three different experimental tasks, Bardovi-Harlig (1998) investigates tense-aspect morphology in interlanguage using a corpus of spoken and written narratives elicited by means of a film-retell task; Waara (2004) compares data from corpora of test interviews of pairs of students (in English as their L1 and L2) by teachers using corpus-linguistic methods and software; similar examples abound in the literature. Similarly, in ways reminiscent of the analysis of parallel corpora, Cadierno (2004, this volume) compares Spanish and Danish spoken narratives produced by native speakers and non-native speakers of Spanish who narrate the events unfolding in a picture book story.

4 This is, of course, a slight simplification because, especially in the domain of curriculum planning, what is of interest may be only the patterning in a particular register rather than the language as a whole. However, this does not affect the general argument here.

5 For example, in the British National Corpus World edition, the 100 most frequent word forms (i.e., about 0.015% of all types) account for a staggering 47% of all the tokens. This is also true of forms of word classes. For example, George (1963) showed that the morphological verb forms taught in a typical first-year English course account for only 20% of all verb form tokens whereas if the course covered only the seven most frequent morphological forms, these would account for almost 60% of all verb form tokens in his corpus.

6 I will not be concerned with collocational frameworks here. Collocational frameworks are structures of two usually closed-class words surrounding an open-class word slot; examples include *a* N *of* or *be* ADJ *to*; cf. Renouf and Sinclair (1991) or Butler (1998) for discussion and application.

7 It is worth pointing out that *colligation* is used very differently by different scholars. Originally coined by Firth, it referred to co-occurrences of grammatical elements, i.e., by analogy to collocations as co-occurrences of lexical elements. The majority of corpus linguists seems to use the term as I do here, namely to refer to co-occurrences of lexical elements with grammatical elements. More recently, Hoey (e.g., Hoey 2003) has broadened the scope of

colligation considerably to also include the preference of lexical elements to particular positions in texts, etc. It remains to be seen whether this extension will be accepted in the field. The notion of *collostruction* basically corresponds to the second definition of colligation but, as the blend of *collocation* and *construction* already suggests, is more strongly related to a Construction Grammar approach and comes with a statistically more sophisticated approach than previous work on Pattern Grammar or colligations (cf. Stefanowitsch and Gries 2003 as well as Gries, Hampe, and Schönefeld 2005).

8 The programs reviewed in Wiechmann and Fuhs (2006) are MonoConc Pro 2.2, WordSmith Tools 4, Concordance, Multi-Language Corpus Tool, ConcApp 4, AntConc 1.3, Aconcorde, Simple Concordance Program, Concordancer for Windows 2.0, and TextStat 2.6. Links to these programs can be found on the author's website.

9 This may seem a trivial point but in fact this is more important than one might think. For example, in the study mentioned above, Waara (2004) states "[e]very occurrence of *get* was extracted from the corpora, i.e., *get, got, getting, gotten*, and *to get*," and one wonders whether the fact that *gets* is missing from this list is responsible for the difference of 157 instances in Tables 1 and 2 of that paper.

10 For example, MonoConc Pro 2.2's (build 242) frequency lists of files from the British National Corpus World edition includes all tags and all words from the header in the frequency list even if the makeup of the tags is specified.

## Bibliography


Aarts, J. and S. Granger. 1998. Tag sequences in learner corpora: a key to interlanguage grammar and discourse. In: Granger, S. (ed.). *Learner English on computer*. London and New York: Addison-Wesley. pp. 132–41.

Altenberg, B. and S. Granger. 2001. The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied Linguistics* 22.2:173–95.

Baayen, R. H. 2001. *Word frequency distributions*. Dordrecht: Kluwer.

Baugh, S., A. Harley, and S. Jellis. 1996. The role of corpora in compiling the Cambridge Dictionary of English. *International Journal of Corpus Linguistics* 1.1:39–59.

Bencini, G. and A. E. Goldberg. 2000. The contribution of argument structure constructions to sentence meaning. *Journal of Memory and Language* 43.4:640–51.

Biber, D., S. Johansson, G. Leech, S. Conrad, and E. Finnegan. 1999. *Longman grammar of spoken and written English*. Harlow: Pearson Education Limited.

Biber, D. and R. Reppen. 2002. What does frequency have to do with grammar teaching? *Studies in Second Language Acquisition* 24.2:199–208.

Bley-Vroman, R. 2002. Frequency in production, comprehension, and acquisition. *Studies in Second Language Acquisition* 24.2:209–13.

Borin, L. and K. Pütz. 2004. New wine in old skins? A corpus investigation of L1 syntactic transfer in learner language. In: Aston, G., S. Bernardini, and D. Stewart (eds.). *Corpora and language learners*. Amsterdam and Philadelphia: John Benjamins, pp. 67–87.

Butler, C. S. 1998. Collocational frameworks in Spanish. *International Journal of Corpus Linguistics* 3.1:1–32.

Bybee, J. and J. Scheibman. 1999. The effect of usage on degrees of constituency: The reduction of *don't* in English. *Linguistics* 37.4:575–96.

Cadierno, T. 2004. Expressing motion events in a second language. In: Achard, M. and S. Niemeier (eds.). *Cognitive linguistics, second language acquisition, and foreign language teaching*. Berlin and New York: Mouton de Gruyter, pp. 13–49.

Chipere, N., D. Malvern, and B. Richards. 2004. Using a corpus of children's writing to test a solution to the same size problem affecting type-token ratios. In: Aston, G., S. Bernardini, and D. Stewart (eds.). *Corpora and language learners*. Amsterdam and Philadelphia: John Benjamins, pp. 137–47.

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.

De Cock, S., S. Granger, G. Leech and T. McEnery. 1998. An automated approach to the phrasicon of EFL learners. In: Granger, S. (ed.). *Learner English on computer*. London and New York: Addison Wesley Longman, pp. 67–79.

Ellis, N. C. 2002a. Frequency effects in language processing and acquisition. *Studies in Second Language Acquisition* 24.2:143–88.

Ellis, N. C. 2002b. Reflections on frequency effects in language processing. *Studies in Second Language Acquisition* 24.2:297–339.

Ellis, N. C. 2006. Language acquisition as rational contingency learning. *Applied Linguistics* 27.1:1–24.

Ellis, N. C. and R. Schmidt. 1997. Morphology and longer distance dependencies. *Studies in Second Language Acquisition* 19.2:145–71.

Eubank, L. and K. R. Gregg. 2002. News flash—Hume still dead. *Studies in Second Language Acquisition* 24.2:237–47.

Evert, S. and B. Krenn. 2001. Methods for the qualitative evaluation of lexical association measures. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pp. 188–95.

Gabrielatos, C. 2005. Corpora and language teaching: just a fling or wedding bells? *Teaching English as a Second or Foreign Language* 8.4: A1.

Gass, S. and L. Selinker. 2001. *Second Language Acquisition: An Introductory Course*. 2nd ed. Mahwah, NJ: Lawrence Erlbaum.

George, H.V. 1963. *A verb-form frequency count: application to course design*. Hyderabad: Central Institute of English.

Glisan, E. W. and V. Drescher. 1993. Textbook grammar: Does it reflect native speaker speech. *The Modern Language Journal* 77.1:23–33.

Grabowski, E. and D. Mindt. 1995. A corpus-based learning list of irregular verbs in English. *International Computer Archive of Modern English (ICAME) Journal* 19:5–22.

Granger S., E. Dagneaux, and F. Meunier. 2002. *The International Corpus of Learner English. Handbook and CD-ROM*. Louvain la Neuve: Presses Universitaires de Louvain.

Granger, S. and M. Wynne. 1999. Optimising measures of lexical variation in EFL learner corpora. In: Kirk, J. (ed.). *Corpora galore*. Amsterdam and Atlanta: Rodopi, pp. 249–57.

Granger, S. 2002. A bird's eye view of learner corpus research. In: Granger, S., J. Hung, and S. Petch-Tyson (eds.). *Computer learner corpora, second language*



*acquisition and foreign language teaching*. Amsterdam/Philadelphia: Benjamins, pp. 3–33.

Gries, S. Th. 2006. Corpus-based methods and cognitive semantics: The many meanings of *to run*. In: Gries, S. Th. and A. Stefanowitsch (eds.). *Corpora in Cognitive Linguistics: Corpus-Based Approaches to Syntax and Lexis*. Berlin, Heidelberg, New York: Mouton de Gruyter, pp. 57–99.

Gries, S. Th. to appear a. Exploring variability within and between corpora: some methodological considerations. *Corpora*.

Gries, S. Th. to appear b. Phraseology and linguistic theory: a brief survey. In: Granger, S. and F. Meunier (eds.). *Phraseology: an interdisciplinary perspective*. Amsterdam and Philadelphia: John Benjamins.

Gries, S. Th. and D. Divjak. submitted. Behavioral profiles: a corpus-based approach to cognitive semantic analysis.

Gries, S. Th., B. Hampe, and D. Schönefeld. 2005. Converging evidence: bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics* 16.4:635–76.

Gries, S. Th., B. Hampe, & D. Schönefeld. to appear. Converging evidence II: more on the association of verbs and constructions. In: Newman, J. and S. Rice (eds.). *Empirical and Experimental Methods in Cognitive/Functional Research*. Stanford, CA: CSLI.

Gries, S. Th. and S. Wulff. 2005. Do foreign language learners also have constructions? Evidence from priming, sorting, and corpora. *Annual Review of Cognitive Linguistics* 3:182–200.

Harrington, M. and S. Dennis. 2002. Input-driven language learning. *Studies in Second Language Acquisition* 24.2:261–8.

Hazenberg, S. and J. H. Hulstijn. 1996. Defining a minimal receptive second-language vocabulary for non-native university students: an empirical investigation. *Applied Linguistics* 17.2:145–63.

Hoey, M. 2000. A world beyond collocation: new perspectives on vocabulary teaching. In: Lewis, M. (ed.) *Teaching collocations*. Hove, UK: Language Teaching Publications, 224–43.

Holmes, J. 1988. Doubt and certainty in ESL textbooks. *Applied Linguistics* 9.1:21–44.

Howarth, P. 1998. Phraseology and second language proficiency. *Applied Linguistics* 19.1:24–44.

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

Hunston, S. and G. Francis. 1999. *Pattern grammar: a corpus-driven approach to the lexical grammar of English*. Amsterdam and Philadelphia: John Benjamins.

Juffs, A. 2001. Verb classes, event structure, and second language learners' knowledge of semantics-syntax correspondences. *Studies in Second Language Acquisition* 23.2:305–13.

Jurafsky, D., A. Bell, M. Gregory, and W. D. Raymond. 2001. In: Bybee, J. and P. Hopper (eds.). *Frequency and the emergence of linguistic structure*. Amsterdam and Philadelphia: John Benjamins, pp. 229–54.

Kempe, V. and B. MacWhinney. 1998. The acquisition of case marking by adult learners of Russian and German. *Studies in Second Language Acquisition* 20.4:534–87.







Kennedy, G. D. 1987. Quantification and the use of English: a case study of one aspect of the learner's task. *Applied Linguistics* 8.3:264–86.

Kita, K., Y. Kato, T. Omoto, and Y. Yano. 1994. A comparative study of automatic extraction of collocations from corpora: Mutual information vs. cost criteria. *Journal of Natural Language Processing* 1.1:21–33.

Laufer, B. and P. Nation. 1995. Vocabulary size and use: lexical richness in L2 written production. *Applied Linguistics* 16.3: 307–22.

Liang, J. 2002. *How do Chinese EFL learners construction sentence meaning: Verb-centered or construction-based?* M.A. thesis, Guangdong University of Foreign Studies.

Ljung, M. 1990. *A study of TEFL vocabulary*. Stockholm: Almqvist and Wiksell.

Mason, O. 1997. The weight of words: an investigation of lexical gravity. *Proceedings of PALC* 1997, pp. 361–75.

Mason, O. 1999. Parameters of collocation: the word in the centre of gravity. In: Kirk, J. (ed.). *Corpora galore: analyses and techniques in describing English*. Amsterdam and Atlanta: Rodopi, pp. 267–80.

Mauranen, A. 2004. Speech corpora in the classroom. In: Aston, G., S. Bernardini, and D. Stewart (eds.). *Corpora and language learners*. Amsterdam and Philadelphia: John Benjamins, pp. 195–211.

McEnery, T. and A. Wilson. 1996. *Corpus linguistics*. 1st ed. Edinburgh: Edinburgh University Press.

Meara, P. 2005. Lexical frequency profiles: a Monte Carlo analysis. *Applied Linguistics* 26.1:32–47.

Mindt, D. 1995. *An empirical grammar of the English verb: modal verbs*. Berlin: Cornelsen.

Mitchell, R. and F. Myles. 1998. *Second language learning theories*. London and others: Arnold.

Nesselhauf, N. 2004. How learner corpus analysis can contribute to language teaching. In: Aston, G., S. Bernardini, and D. Stewart (eds.). *Corpora and language learners*. Amsterdam and Philadelphia: John Benjamins, pp. 109–24.

Read, J. 1997. Assessing vocabulary in a second language. In: Clapham, C. and D. Corson (eds.). *Language testing and assessment. Encyclopedia of language and education*, Vol. 7. Dordrecht: Kluwer, pp. 99–107.

Read, J. 2000. *Assessing vocabulary*. Cambridge: Cambridge University Press.

Renouf, A. and J. M. Sinclair. 1991. Collocational frameworks in English. In: Aijmer, K. and B. Altenberg (eds.). *English corpus linguistics*. London: Longman, pp. 128–43.

Römer, U. 2004. Comparing real and ideal language learner input. In: Aston, G., S. Bernardini, and D. Stewart (eds.). *Corpora and language learners*. Amsterdam and Philadelphia: John Benjamins, pp. 151–68.

Römer, U. 2005. *Progressive, patterns, pedagogy*. Amsterdam and Philadelphia: John Benjamins.

Schönefeld, D. 1999. Corpus linguistics and cognitivism. *International Journal of Corpus Linguistics* 4.1:131–71.

Stefanowitsch, A. and S. Th. Gries. 2003. Collostructions: investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8.2:209–43.

Tono, Y. 2002. The role of learner corpora in SLA research and foreign language





teaching: the multiple comparison approach. Unpublished Ph.D. thesis, University of Lancaster.

Tono, Y. 2004. Multiple comparisons of IL, L1 and TL corpora: the case of L2 acquisition of verb subcategorization patterns by Japanese learners of English. In: Aston, G., S. Bernardini, and D. Stewart (eds.). *Corpora and language learners*. Amsterdam and Philadelphia: John Benjamins, pp. 45–66.

Waara, Renee. 2004. Construal, convention, and constructions in L2 speech. In: Achard, M. and S. Niemeier (eds.). *Cognitive linguistics, second language acquisition, and foreign language teaching*. Berlin and New York: Mouton de Gruyter, pp. 51–75.

Wiechmann, D. and S. Fuhs. 2006. Concordancing software. *Corpus Linguistics and Linguistic Theory* 2.1:109–30.

Willis, D. 1990. *The lexical syllabus: a new approach to language teaching*. London: Collins Cobuild.

Wolter, B. 2001. Comparing the L1 and L2 mental lexicon. *Studies in Second Language Acquisition* 23.1:41–69.

Xiao, R. and T. McEnery. 2006. Collocation, semantic prosody, and near synonymy: a cross-linguistic perspective. *Applied Linguistics* 27.1:103–29.