4.1 Corpus and Quantitative Methods

Stefan Th. Gries

Chapter Overview

> The behavior of the speaker, listener, and learner of language constitutes, of course, the actual data for any study of language. Chomsky (1959: 59)

1 Introduction

The core question at the heart of nearly all work in cognitive/usage-based linguistics is, how do characteristics of the cognitive system affect, or at least correlate with, the acquisition, representation, processing, use and change of language? Thus, ever since Lakoff's (1990: 40) formulation of the cognitive commitment – the 'commitment to providing a characterization of general principles for language that accords with what is known about the mind and brain from other disciplines' – cognitive/usage-based approaches have revolved around notions such as:

- exemplars and entrenchment;
- chunking and learning;

- association and contingency;
- categorization, prototypicality and schematicity, as well as cue and category validity;
- productivity and creativity;
- analogy and similarity.

Even though these notions all involve human cognition and have been addressed with quite some empirical rigour in, say, psychology or psycholinguistics, the first wave of cognitive-linguistic research was largely and explicitly based on introspection just as the generative approach against which much of Cognitive Linguistics was arguing. For example, the early network analyses of highly polysemous words (most notoriously, *over*) liberally used the language of mental networks but came with little to no empirical data, and introspection or speculation was defended as a necessary element of cognitive-linguistic analysis (e.g. Langacker, 1987 or Talmy, 2000 or recent statements at the retrospective panel of the ICLC, 2013).

However, in the last 20–25 years or so, there has been a greater recognition of the problems that arise when linguists provide both the theory and the data. With regard to polysemy networks, for instance, Sandra and Rice (1995) has been a wake-up call in how they discuss both corpus-linguistic and experimental ways (combined with statistical analyses) to put the study of polysemy networks etc. on firmer empirical grounds. Nowadays, cognitive/usage-based linguistics is characterized by a more widespread adoption of corpus data as a source of relevant linguistic data and quantitative/statistical tools as one of the central methodologies, and the field is now brimming with new corpus-based methods and statistical tools (cf. Ellis, 2012 for a recent comprehensive overview). This chapter will provide a brief overview of how corpus data and statistical methods are used in increasingly sophisticated ways in Cognitive Linguistics. While Cognitive Linguistics does not make a principled distinction between syntax and lexis anymore but rather assumes a syntax-lexis continuum, for expository reasons I will discuss (more) lexical examples in Section 2, (more) syntactic examples in Section 3, and I will then turn to selected applications of quantitative corpus linguistics in phonology and morphology in Section 4. Section 5 will then conclude with a brief discussion of necessary future developments.

> This last point leads me, with some slight trepidation, to make a comment on our field in general, an informal observation based largely on a number of papers I have read as submissions in recent months. In particular, we seem to be witnessing as well a shift in the way some linguists find and utilize data – many papers now use corpora as their primary data, and many use internet data. (Joseph, 2004: 382)

2 Syntax-lexis, with an Emphasis on Lexis

Given its historical association with dictionary-making, corpus linguistics has always had a strong emphasis on the analysis of lexical items. Concordances – lists of uses of words in their authentic contexts – and collocations – tables of words that are used in slots around a word of interest – have long helped lexicographers to tease apart multiple senses of polysemous words or differences in how near synonymous words are used. Especially for collocations, corpus linguists also increasingly rely on association measures to separate the wheat – frequent co-occurrence that reflects interesting semantic and/or functional characteristics – from the chaff – frequent co-occurrence that reflects little of semantic interest, such as the fact that most nouns co-occur a lot with *the*. A syntactically more informed perspective then also studied colligation, that is, the co-occurrence of words or senses with elements in syntactically defined slots; early examples in Cognitive Linguistics are Schmid (1993), Kishner and Gibbs (1996) on *just*, and Gibbs and Matlock (2001) on *make*. While under-appreciated (and ground-breaking) at the time, these studies were still largely monofactorial in nature: Uses of (senses of) words were annotated for, and cross-tabulated with, co-occurrence patterns, but no real quantitative analyses were conducted on the distributional data thus obtained. The current state of the art is that such multidimensional co-occurrence data are also statistically analysed in multidimensional ways. Gries (2010b) distinguishes two different ways in which analyses can be multidimensional, which will be exemplified in the following two sections.

2.1 Multidimensional$_1$ Approaches: Behavioural Profiles and Cluster Analyses

The first sense of *multidimensional*, multidimensional$_1$, refers to the fact that concordance lines of (senses of) a word are annotated for many different characteristics – morphological, syntactic, semantic, discourse-pragmatic – and all of these dimensions are used in a statistical analysis at the same time, but *separately* from each other. One example for this approach that has become more widely used is the behavioural profile (BP) approach (cf. Gries, 2010b for a detailed overview). In this approach, concordance lines are annotated for many features on many dimensions, and then the senses of polysemous words, or the near synonyms in point, are compared with regard to the percentages with which different features are attested with a sense/word. Consider Figure 4.1.1, which represents this process. The upper part illustrates how, in this case, three concordance lines of the verb lemmas *begin* and *start* are annotated for a variety of features. For example, the first concordance line was a line where *begin* was

used in the progressive (*ing*) and the entity that is beginning something was something abstract; the same is done for other concordance lines and for many other features. The lower part of Figure 4.1.1 is then the result of cross-tabulating the frequencies with which types of features are attested with the two lemmas. For instance, 20 per cent of all instances of the lemma *begin* were in the progressive, and 40 per cent of all instances of the lemma *start* were in the progressive, which means the two lemmas are rather different on that dimension. On the other hand, they are quite similar with regard to their past tense use: *begin* and *start* are used in the past tense 40 per cent and 38 per cent respectively. It is the columns of the lower part of Figure 4.1.1 that are referred to as behavioural profiles, since they summarize the percentages with which a lemma is used with/in something else.

Gries (2006) applied this method to the many senses of *to run*, Divjak (2006) studied Russian verbs meaning 'to intend', and both find that the percentages of co-occurrence phenomena reliably distinguish senses and near synonyms respectively. In addition, Gries (2006) also showed how co-occurrence percentages can be used to study the similarity of senses, their positions in networks, whether to lump or split them, and how more generally different types and aspects of corpus data help identify the prototypical senses of words (viz. type and token frequencies, earliest historical attestations, earliest language acquisition attestations, etc.).

A variety of more complex follow-up approaches to BP analyses have been pursued, too. For example, the behavioural profiles of, say, near synonyms with linguistic patterns in their contexts can be submitted to exploratory statistical tools such as hierarchical cluster analyses. Divjak and Gries (2006) is a case in point. They studied nine Russian verbs meaning 'to try' and analyse the similarity of BP co-occurrence percentages with cluster analyses and follow-up exploration in terms of average silhouette widths, *t*- and *F*-scores, etc. They found that this lexical field falls into three different groups (of three verbs each), which reflect different idealized cognitive models of trying. Even more interestingly, though, is that Divjak and Gries (2008) showed that the clusters obtained on the basis of the corpus analysis are very strongly replicated in sorting and gap-filling experiments with native speakers of Russian, a finding that testifies to the reliability and validity of the BP approach. Finally, Janda and Solovyev (2009) used a downsized version of BP data – the constructional profile, the relative frequency distribution of the grammatical constructions a word occurs in – to explore synonyms.

A final BP example to be mentioned showcases the potential of the BP approach for cross-linguistic analysis. Divjak and Gries (2009) studied phasal verbs in English (*begin* vs *start*) and Russian (*načinat'/načat'*, *načinat'sja/načat'sja*, and *stat'*). Computing, among other things, pairwise differences between behavioural profiles – as discussed above for progressive and past

Concordance line	Verb lemma	Verb form	What begins	...
1	*begin*	ing	abstract	...
2	*start*	past	human	...
3	*start*	infin	human	...
...	...	...	...	...

↓↓↓↓↓

ID tag	ID tag level	*begin*	*start*
Verb form	ing	0.2	0.4
	past	0.4	0.38
	infin	0.3	0.1
	...	...	...
What begins	abstract	0.15	0.2
	human	0.4	0.2
	...	...	...
...	...	...	...

Figure 4.1.1 Schematic representation of a BP analysis (fictitious numbers)

tense uses of *begin* and *start*, within English, they found that *start* is more frequent than *begin* with scenarios where human instigators start (esp. communicative) actions, and within Russian, *načinat'/načat'* prefers imperfective aspect and situations with a clear beginning whereas *stat'* prefers perfective aspect and actions instigated by humans. This is represented in dotcharts in Figure 4.1.2 and Figure 4.1.3: the percentage differences between the verbs being compared are on the *x*-axis, the differences are sorted by features and then by size, and the three vertical lines indicate the mean of all differences and its confidence interval. Thus, differences outside of this interval can be easily identified and point to potentially interesting distributional differences of the verbs.

However, since the annotated features are cross-linguistically comparable, Divjak and Gries also compared specific English to Russian verbs and, more generally, explored the features that make English speakers choose one of the synonyms as compared to Russian speakers. For instance, they found that English speakers' choices are driven by semantic characteristics of the beginners and beginnees whereas Russian speakers' choices are driven by aspectual and argument-structural characteristics.

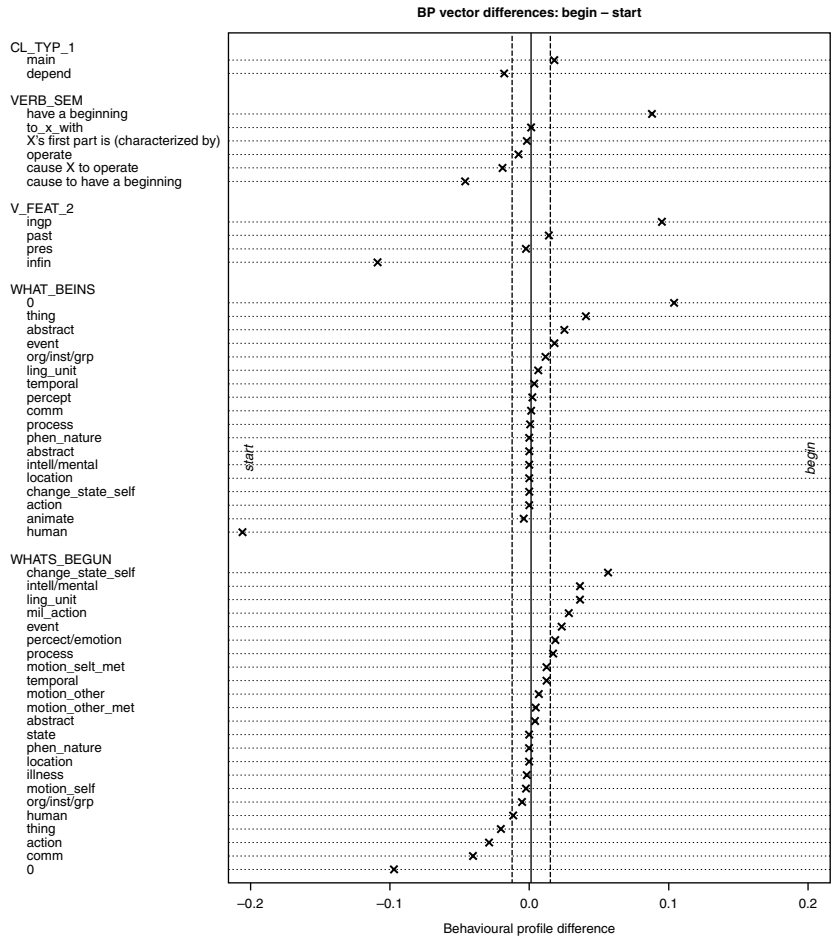

Figure 4.1.2 BP vector preferences contrasting *begin* and *start* (based on data from Divjak and Gries, 2009)

2.2 Multidimensional$_2$ Approaches: Regression and Correspondence Analysis

The second sense of *multidimensional*, multidimensional$_2$, refers to the fact that concordance lines of (senses of) a word are annotated for many different characteristics – morphological, syntactic, semantic, discourse-pragmatic – and all of these dimensions are used in a statistical analysis *together.* That is, multidimensional$_1$ uses the information of how a linguistic item – a morpheme, a word, a sense, . . . – behaves on each of many dimensions such as

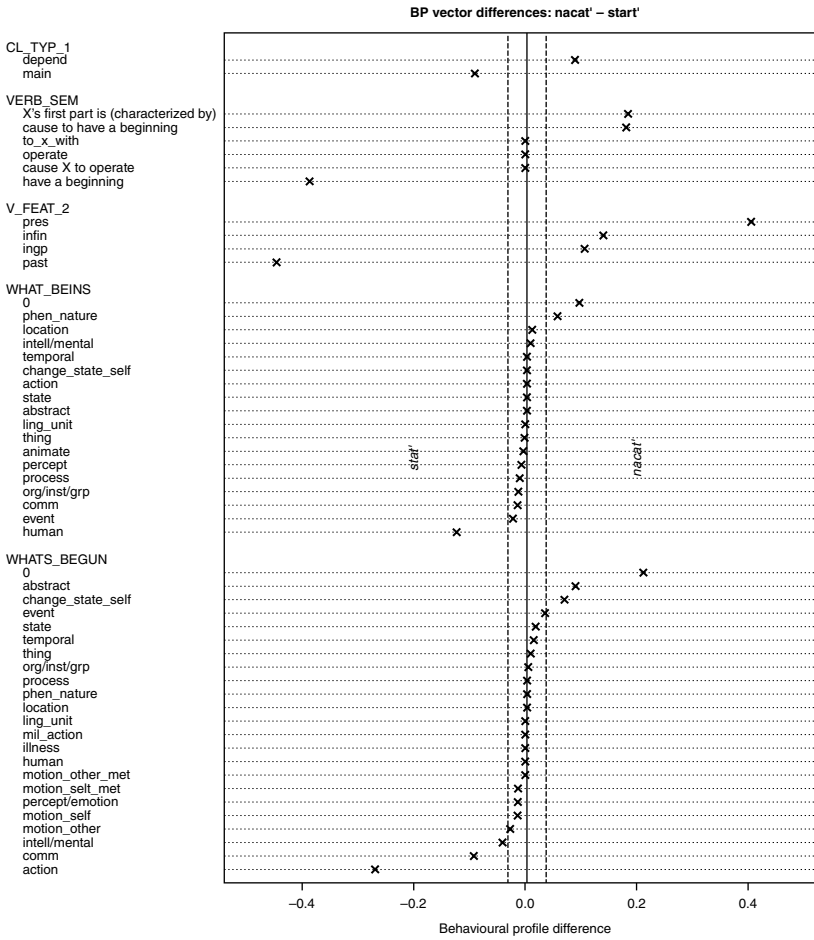

Figure 4.1.3 BP vector preferences contrasting *nacat'* and *stat'* (based on data from Divjak and Gries, 2009)

- what are the percentages with which sense *x* has different kinds of subjects?
- what are the percentages with which sense *x* has different kinds of objects? etc.

For example, if one annotates *n*=2 dimensions of variation – for example, the percentages of different subjects of senses *a* to *f* and the percentages of different objects of senses *a* to *f* – then multidimensional$_1$ analysis uses that information in the shape of combining results from *n*=2 two-dimensional frequency/percentage tables. But what is not included are the co-occurrence percentages of sense *x*'s different subjects *with its different objects* – this is what multidimensional$_2$

does in the shape of one three-dimensional table: sense (*a* to *f*) × subject (all subject types) × object (all object types). The advantage over the BP analysis is, therefore, that higher-level co-occurrence information is included, which is more precise and cognitively more realistic (although, recall the strong experimental validation of the BP approach). The disadvantage is that this can easily lead to very sparse data sets, as when many features are annotated so that any actual combination of features is very rare.

Two types of multidimensional$_2$ applications are particularly interesting. First, exploratory approaches such as those using (multiple) correspondence analysis (MCA), a method applied to multidimensional frequency data that is similar to principal component analysis. One such application to a polysemous word is Glynn's (2010) study of *bother*. Glynn followed the work discussed in Section 2.1 and annotated uses of *bother* for a large number of features and applied MCAs to different parts of the multidimensional frequency table. The results revealed different clusters and 'semantically motivated distinction[s] between two sets of syntactic patterns' (Glynn, 2010: 256), an agentive and a predicative construction. In order to test the patterns suggested by the exploratory tool, Glynn then added the second type of multidimensional$_2$ application, confirmatory approaches based on regression analyses. In this case, he ran a binary logistic regression to determine to what extent the co-occurrence features of *bother* distinguish between the two constructions. His analysis resulted in a good classification accuracy, showing that, just like BPs, a careful multidimensional analysis of corpus data with powerful statistical tools can reveal cognitively and constructionally interesting regularities impossible to discover by intuition or eyeballing of data. Additional applications of this approach in the domain of semantics include Glynn (2012), a replication of Gries (2006) and, with a fascinating interpretation of the notion of *corpus*, Levshina's (in prep.) study of how an MCA discovers structure in the semantic field of seating furniture, where the different words for pieces of furniture are annotated for characteristics taken from German online furniture catalogues such as 'ab-/presence of armrests', 'use of upholstery', 'back recline', 'seat surface recline', etc.

Additional examples for similar multidimensional$_2$ applications involve binary as well as multinomial or polytomous logistic regressions. As for the former, Deshor and Gries (2012) compared the uses of *may* and *can* by native speakers of English and French to see how well syntactic and semantic features allow to predict speakers' choices, but also to determine which variables distinguish the native speaker's from the learners' use of *may* and *can*; the results were then interpreted against the background of processing principles. As for the latter, Arppe (2008) studied four common Finnish verbs meaning 'to think' by, as usual, annotating them for a variety of linguistic characteristics and then identifying the linguistic characteristics that best allow to predict speakers' choices; later work by Divjak and Arppe (e.g. 2010) extended such regression

approaches to the identification of prototypes in a way inspired by, but not referencing, Gries (2003b), who uses linear discriminant analysis to the same end (a classifier mathematically different from, but nonetheless comparable to, the now common regression models).

Regardless of which multidimensional approach is chosen, the combination of comprehensive annotation and multifactorial/-variate analysis has yielded insightful results regarding a variety of the above-mentioned central notions of Cognitive Linguistics on the level of lexical items, including the degree to which words/senses are entrenched, the association/contingency of formal and functional elements, matters of categorization (graded similarity vs discreteness of senses, prototypes of senses) and many more. For more examples regarding the corpus-based exploration of metaphor and metonymy, the reader is referred to the collection of papers in Stefanowitsch and Gries (2006); for more examples highlighting in particular statistical applications, see Glynn and Fischer (2010) and Glynn and Robinson (2012). The following section will now turn to the more syntactic side of the syntax-lexis continuum.

> Linguistics has always had a numerical and mathematical side [. . .], but the use of quantitative methods, and, relatedly, formalizations and modeling, seems to be ever on the increase; rare is the paper that does not report on some statistical analysis of relevant data or offer some model of the problem at hand. (Joseph, 2008: 687)

3 Syntax-lexis, with an Emphasis on Syntax

Not unsurprisingly, the corpus-linguistic tools used on the more syntactic side of the continuum are quite similar to those on the more lexical side of things. Again, concordances are used to explore the use of syntactic patterns, or constructions, in their context, and colligations/collexemes – tables of words occurring in syntactically defined slots of constructions – are used to explore the ways in which constructional slots are filled. One major difference of course is concerned with the searchability of constructions, since corpora that are annotated for constructions in the general sense of the term do not exist. Thus, corpus searches for constructions typically rely on words (searching for `way` or `into [a-z]ing` to find the *way* construction or the *into*-causative), part of speech tags (searching for `DPS way` [`DPS` = possessive determiner] or `into VVG` [`VVG` = lexical verb in the progressive] to find the *way* construction or the *into*-causative), parsed corpora, or combinations of all these things with lots of subsequent manual disambiguation. In the following sections, I will first discuss a recent development in the study of colligations/collexemes, which is a simple monofactorial topic, before I turn to corpus-linguistically and quantitatively more involved topics.

3.1 Monofactorial Approaches: Frequencies, Percentages and Collostructions

One recent prominent approach in the study of constructions – the way they fill their slots and what that reveals about their semantics/function – is collostructional analysis (Stefanowitsch and Gries, 2003, 2005 and Gries and Stefanowitsch, 2004a, b). By analogy to collocations, Gries and Stefanowitsch proposed to study the functions of constructions by not just looking at how frequently words occur in their slots (e.g. which verbs occur in the verb slot of the *way* construction how often?) but by computing measures of association (most often $p_{\text{Fisher-Yates exact test}}$) that quantify how strongly (or weakly) a word and a construction are attracted to, or repelled by, each other. This family of methods has some psycholinguistic foundation and has been widely adopted in studies on near-synonymous constructions (alternations), priming effects (Szmrecsanyi, 2006), first and second language acquisition and learning of constructions (cf. Ellis and Ferreira-Junior, 2009; Goldberg, 2006 and Gries and Wulff, 2005, 2009 for insightful discussion of many compatible findings), constructional change over time (Hilpert, 2006, 2008), etc. For alternations, for instance, the method was precise enough to discover the iconicity difference between the ditransitive (small distances between recipient and patient) and the prepositional dative (larger distances between recipient and patient; cf. Thompson and Koide, 1987).

In the last few years, a variety of studies have been published which also document the validity of the method experimentally. Gries, Hampe and Schönefeld (2005) demonstrated how collexeme analysis outperforms frequency and conditional probabilities as predictors of subjects' behaviour in a sentence completion task, and the follow-up of Gries, Hampe and Schönefeld (2005) provided additional support from self-paced reading times; cf. also Gries (2012) for a comprehensive overview and rebuttal of Bybee (2010: Section 5.12). Lastly, collostructions have been coupled with more advanced statistical tools – such as cluster analysis or correspondence analysis – to discover sub-senses of constructions (cf. Gries and Stefanowitsch, 2010) or structure in lexical fields (when this tool is applied to lexical items, cf. Desagulier, 2012).

3.2 Multidimensional$_2$ Approaches: Regression and Correspondence Analysis

The previous section already mentioned the use of advanced statistical tools in the analysis of constructions; in the terminology of Section 2, these tools are multidimensional$_2$ and I will again discuss examples using exploratory and confirmatory approaches; for expository (and historical) reasons, I will begin with the latter.

As far as I can see, the first multifactorial approaches in cognitive corpus linguistics were Gries's (2000, 2003a) studies of the constructional alternation of particle placement, that is the two constructions instantiated by *Picard picked up the tricorder* and *Picard picked the tricorder up*. On the basis of corpus data from the British National Corpus (BNC), he annotated examples of both constructions for a large number of phonological, morphological, syntactic, semantic, and discourse-functional parameters and used a linear discriminant analysis to identify the factors that make speakers choose one construction over another in a particular discourse context, discuss their implications for language production and identify prototypical instances of both constructions. Since then, this type of approach – multifactorial modelling syntactic, but now also lexical, alternatives with regression-like methods – has become very prominent both within and outside of cognitive linguistics proper and within Cognitive Linguistics there are at least some studies that show how well this approach helps explore such alternations; Szmrecsanyi (2006), Gries and Wulff (2009) are two cases in point using logistic regressions, Levshina, Geeraerts, Speelman (2012) for the additional tool of classification and regression trees, and Gries (2003b) showed that the predictions of such methods correlate very strongly with results from acceptability ratings.

There are also exploratory approaches to be discussed, and again they involve the method of multiple correspondence analysis. One particularly interesting example involves the cross-linguistic corpus-based study of analytic causatives in English and Dutch. On the basis of data from the newspaper component of the BNC (approx. 10m words) for English and an equally large sample from the Twente and the Leuven News corpora, Levshina, Geeraerts and Speelman (2013) retrieved approx. 4,000 examples of causatives from both languages, which were annotated for the semantic classes of the causer and the causee as well as for one of many different semantic verb classes. An MCA was then used to determine the conceptual space of the causatives in the two languages. Among other things, this bottom-up procedure provided a two-dimensional representation (of an ultimately three-dimensional) conceptual causative space with clear support for a previous merely theoretical typology of causative events. In addition, a follow-up analysis of the results of separate analyses of the English and the Dutch data showed that the two languages' conceptual causative space is overall similar, but not identical, and the authors discussed how both languages' data points are located differently in causative space.

3.3 Straddling the Boundaries of Lexis and Syntax: Idioms and Multiword Units

As mentioned above and for purely expository reasons, Sections 2 and 3 in this chapter upheld a distinction that cognitive linguists – and many corpus

linguists – do not usually make anymore, the one between syntax and lexis. In fact, many of the earliest studies in Construction Grammar focused on items straddling the 'syntax-lexis boundary', namely constructions that were traditionally called idioms (cf. Wulff, 2008 for the probably most rigorous cognitive and corpus-linguistic study of idiomaticity). At that time, and in fact until recently, it was part of the definition of construction that an expression considered a candidate for constructionhood exhibited something that was not predictable from its constituent parts and other constructions already postulated. While, in Goldbergian Construction Grammar, this notion of unpredictability is no longer a necessary condition, there is now also a growing body of research on the psycholinguistic status multiword units (MWUs, also often called *lexical bundles*), that is expressions consisting of several contiguous words. On the one hand, MWUs do not seem good candidates for constructionhood since they are often not even 'proper' phrasal elements, do not have a particularly unified semantic/functional pole, and have little that is unpredictable about them, but on the other hand many of them, at some point, became retained in speakers' minds and, thus, most likely also gave rise to processes of chunking (cf. Bybee, 2010: Ch. 3, 8). Many such studies are experimental in nature but usually take their starting point from corpus frequencies of MWUs. For instance, Bod (2000) showed that high-frequency 3-grams (e.g. *I like it*) are reacted to faster than lower-frequency 3-grams (e.g. *I keep it*), and Lemke, Tremblay and Tucker (2009) provided evidence from lab-induced speech that the last word of a 4-gram is more predictable than expected by chance, which they interpreted as showing that MWUs are stored as lexical units; similar findings are reported by Huang, Wible and Ko (2012) based on the comparison of transitional probabilities in corpus data and eye-tracking data; cf. for more discussion Arnon and Snider (2010), Snider and Arnon (2012), and Caldwell-Harris, Berant and Edelman (2012).

Again, the analysis of many of the central notions of the cognitive/usage-based approach to language benefits in multiple ways from the combination of fine-grained annotation of corpus data and powerful statistical tools, which elucidate complex patterns and interactions in the data that defy introspective or simple monofactorial analysis: notions such as chunking and entrenchment of words into MWUs, association and contingency of words in constructional slots (which are based on the validity of cues and constructional categories), the implications of this for learnability and processing . . . all these are areas where state-of-the-art quantitative corpus linguistics can be very useful. For more examples, see Stefanowitsch (2010a) and the papers in Gries and Stefanowitsch (2006), Rice and Newman (2010), Schönefeld (2011), Divjak and Gries (2012), and Gries and Divjak (2012).

> Now that corpus linguistics is turning more and more into an integral part of mainstream linguistics, [. . .] we have to face the challenge of complementing

> the amazing efforts in the compilation, annotation and analysis of corpora with a powerful statistical toolkit and of making intelligent use of the quantitative methods that are available. (Mukherjee, 2007: 141)

4 Phonology and Morphology

For purely technological reasons, corpus linguistics has been particularly involved in studies on lexis and syntax. However, given increasingly more and larger resources as well as the ongoing development of new techniques and tools, there is now also a considerable body of corpus-based cognitive-linguistic research in domains such as phonology and morphology. Space does not permit an exhaustive discussion but the following sections highlight some examples.

4.1 Phonology

Some of the more influential recent studies on phonological reduction were not cognitive-linguistic in a narrower sense, but certainly compatible with current cognitive-linguistic work on processing. As one example, Bell et al. (2003) is a comprehensive study using regression analyses on how the pronunciation of monosyllabic function words (in the Switchboard corpus) is affected by disfluencies, contextual predictability (measured in terms of transitional probabilities, and earlier studies used the association measure *MI*), and utterance position.

To mention one more recent example, Raymond and Brown (2012) used binary logistic regression to study initial-fricative reduction in Spanish. Their study is remarkable for the range of variables they take into consideration to shed light on why many studies of frequency effects come to contradictory results. Maybe the most important conclusion is that, once contextual probabilities are taken into account, non-contextual frequencies did not yield any robust results, a finding strongly supporting the view that simple frequencies of occurrence are often not enough.

4.2 Morphology

Another area in which corpus-based studies have had a lot to offer to Cognitive Linguistics is morphology. There is a large number of studies by Bybee and colleagues (nicely summarized in Bybee, 2010) that revealed how frequency of (co-)occurrence affects chunking or resistance to morphosyntactic change, to

name but some examples, and that have been integrated into a usage-based network model of morphology. A different though ultimately related strand of research is work on morphological productivity, specifically on how to measure it best and how relative frequency – the difference in frequency of derived words (e.g. *inaccurate*) and their bases (e.g. *accurate*) – affects productivity as well as morphological processing, which in turn informs theoretical discussions of decompositional vs non-decompositional approaches; cf. Hay and Baayen (2003) or Antić (2012) for a more recent contribution.

Let me finally mention a few smaller case studies. On the basis of a small corpus of Dena'ina narratives, Berez and Gries (2010) explored the factors that trigger the ab-/presence of the middle marker *d* in iterative verbs. Traditionally, *d* was considered a reflex of syntactic transitivity, with semantics playing a less important role. However, a binary logistic regression and a hierarchical configural frequency analysis of their data showed that, while transitivity is a relevant predictor, the semantic type of iterativity (and its position on a scale from concrete to abstract) resulted in an even higher degree of predictive power.

Lastly, Teddiman (2012) showed how subjects' decisions on which part of speech to assign to ambiguous words in an experiment are very strongly correlated (r_S=0.87) with the words' preferences in the CELEX database. On the whole, words such as *pipe* and *drive* (mostly used nominally and verbally respectively) were typically assigned to be nouns and verbs respectively.

4.3 Straddling the Boundaries of Phonology and Morphology

Just as there are phenomena somewhere between, or in both lexis and syntax, so there are phenomena somewhere between, or in both phonology and morphology. An example of the former is Bergen (2004) on phonaesthemes. While the main point of his study involved a priming experiment, one section of it showed how some phonaesthemes such as *gl-*, *sn-*, and *sm-* are significantly more often attested with their phonaesthemic meanings of 'light' and 'nose/mouth' than expected by chance, which raises interesting issues for classical morphological theory, into which phonaesthemes do not fit very well, and statistical learning by speakers.

An example of the latter, a phenomenon 'in' both phonology and morphology is blends, formations such as *motel* (*motor* × *hotel*) or *brunch* (*breakfast* × *lunch*). In a series of studies, Gries showed how coiners of such blends have to strike a balance between different and often conflicting facets of phonological similarity and semantics while at the same time preserving the recognizability of the two source words entering into the blend. Again, this corpus-informed

work sheds light on a phenomenon that traditional morphology finds difficult to cope with.

> We constantly read and hear new sequences of words, recognize them as sentences, and understand them. It is easy to show that the new events that we accept and understand as sentences are not related to those with which we are familiar by any simple notion of formal (or semantic or statistical) similarity or identity of grammatical frame. (Chomsky, 1959: 59)

5 Concluding Remarks and Future Developments

As the previous sections have demonstrated, corpus linguistic methods have become an important component of cognitive/usage-based linguistics. This methodological development seems to have happened in tandem with a shift in linguistics in general, as evidenced by some epigraphs in this chapter, but also with a shift within Cognitive Linguistics, as evidenced by the fact – unthinkable ten years ago – that Mouton just published a reader called *Cognitive Linguistics: The Quantitative Turn* (Janda, 2013). While Cognitive Grammar had a strong commitment to being usage-based ever since Langacker's *Foundations of Cognitive Grammar*, other parts of Cognitive Linguistics – that is, the Lakovian 'branch' of Cognitive Linguistics and/or early Construction Grammar – put much less emphasis on the usage-based nature of grammar/language. Now that the theory of Cognitive Linguistics as a whole has become much more usage-based, it is only fitting that analyses of actual usage – corpus data – play a much more central role. The type of exemplar-based approaches that many cognitive linguists now embrace are particularly compatible with the distributional data that corpora provide, and it is especially in this way that corpus linguistics and cognitive/usage-based linguistics inform each other. For instance, the following are examples of how the theoretical framework of usage-based linguistics relies on, and is advanced and informed by, corpus linguistic tools:

- the overall frequency of elements is a proxy to their entrenchment;
- the degree to which elements are more frequent in combinations with other elements or behave differently from when they are used in isolation informs our thinking of how elements are chunked into units;
- the way in which corpus data allows us to measure predictive co-occurrence allows us to explore the multidimensional exemplar space that, according to usage-based linguists, contains both linguistic and encyclopedic knowledge;

- the way how frequency data from corpora give rise to clusters in multi-dimensional space reflects our views of prototypes (as densely populated regions of space with configurations of highly predictive features, which can often just be cue and category validities directly measured from corpus data; cf. Goldberg, 2006); etc. etc.

At the same time, cognitive/usage-based linguistics provides a much-needed dose of a theoretical framework to corpus linguistics, a field that is still often merely descriptive and even reluctant to embrace (certain more theoretical) generalizations (cf. Gries, 2010c for much discussion).

In these next brief sections, I would like to very briefly provide some comments on where I think Cognitive Linguistics can and should evolve and mature further by incorporating insights from quantitative corpus linguistics.

5.1 More and Better Corpus-linguistic Methods

One important area for future research is concerned with refining the arsenal of corpus-linguistic tools. First, there is a growing recognition of the relevance of association measures in cognitive/usage-based linguistics. However, with very few exceptions, such association measures are bi-directional or symmetric: they quantify the attraction of x and y to each other as opposed to the attraction of x to y, or of y to x, which would often be psychologically/psycholinguistically more realistic. Gries (2013b), following Ellis (2007) and Ellis and Ferreira-Junior (2009), discussed and validated a directional association measure from the associative learning literature on the basis of corpus data, which should be interesting for anybody dealing with association and contingency, say in language learning/acquisition. Similarly, the entropies of the frequencies of linguistic elements are an important element qualifying the effect of type frequencies in corpus data (cf. Gries, 2013a, b), which in turn affects productivity and flexibility/creativity of expressions (cf. Zeschel, 2012 and Zeldes, 2012) as well as their learnability.

Second, there is now also a growing recognition that corpus frequencies of x and y can be highly misleading if the dispersion of x and y in the corpus in question is not also considered: if x and y are equally frequent in a corpus but x occurs in every corpus file whereas y occurs only in a very small section of the corpus, then y's frequency should perhaps be downgraded, and Gries (2008, 2010) discussed ways to measure this as well as first results that indicate that, sometimes, dispersion is a better predictor of experimental results than frequency.

Finally, there will be, and should be, an increase of corpus-based studies that involve at least some validation against experimental data, as in many of the studies from above.

5.2 More and Better Statistical Tools

Another area that is much in flux involves the development of statistical tools. One approach that is gaining ground rapidly is the technique of new regression-like methods. On the one hand, the technique of mixed-effects (or multi-level) modelling is becoming more frequent, since it allows the analyst to handle subject/speaker-specific and, for example, word-specific variation as well as unbalanced data much better than traditional regression tools. On the other hand, new classification tools such as Bayesian network and memory-based learning (cf. Theijssen et al., to appear) with its ability to model causal effects in a way reminiscent of structural equation modelling and naïve discriminative learning (cf. Baayen, 2010) with its higher degree of cognitive realism are becoming important promising new alternatives. Finally, I hope that exploratory/bottom-up techniques will become more frequently used.

5.3 Additional Developments

I would finally like to offer a few more diverse suggestions as to where the field will, and/or should be going. For instance, I expect that the field of usage-based language acquisition will benefit increasingly more from more and better resources and techniques. Corpus-based studies on the development of early syntax using the traceback method (Dąbrowska and Lieven, 2005), for example, showcase the potential for computational corpus-linguistic work. Similarly, in order to study word and construction learning and the role of preemption, corpus data have and will become more and more important (cf. Stefanowitsch, 2011 and Goldberg's 2011 response).

In addition, I think the field can benefit from a greater recognition of individual differences. Studies such as Street and Dąbrowska (2010) or Caldwell-Harris, Berant and Edelman (2012) and others show clearly that the 'native speaker' to which all linguistic theories like to generalize is little more than a convenient fiction, given the huge individual diversity that both corpus and experimental data reveal very clearly (esp. with mixed-effects models).

To wrap up, Stefanowitsch (2010b) discussed cognitive semantics with regard to three steps of the evolution of a discipline from art to science, (i) adopt the protocols/practices of empirical research, (ii) adopt those to the object of research and operationalize theoretical concepts, and (iii) relegate to metaphysics all concepts that resist such operationalization. While this chapter could only provide the briefest of overviews of the impact that corpora and quantitative methods have had Cognitive Linguistics, it is probably fair to say that they are conquering the field by storm in how they facilitate steps (i) and (ii). It is to be hoped that this development/maturation of the field continues as individual

scholars increase their repertoire of corpus and quantitative skills (cf. Gries, 2013a and Gries and Wulff, in progress) and as more and more fruitful connections with neighbouring disciplines – corpus linguistics or psycholinguistics, to name just two examples – provide ever more opportunities for interdisciplinary research.

References

Antić, E. (2012). Relative frequency effects in Russian morphology. In S. Th. Gries and D. S. Divjak (Eds), *Frequency Effects in Language Learning and Processing*. Berlin and New York: Mouton de Gruyter, pp. 83–107.

Arnon, I. and Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language,* 62(1), 67–82.

Arppe, A. (2008). Univariate, bivariate and multivariate methods in corpus-based lexicography – a study of synonymy. PhD dissertation, University of Helsinki.

Baayen, R. Harald (2010). Corpus linguistics and naïve discriminative learning. *Brazilian Journal of Applied Linguistics,* 11(2), 295–328.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M. and Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *Journal of the Acoustical Society of America,* 113(2), 1001–24.

Berez, Andrea L. and Gries, Stefan Th. (2010). Correlates to middle marking in Dena'ina iterative verbs. *International Journal of American Linguistics,* 76(1), 145–65.

Bergen, Benjamin K. (2004). The psychological reality of phonaesthemes. *Language*, 80(2), 290–311.

Bod, R. (2000). The storage vs. computation of three-word sentences. Paper presented at AMLaP-2000.

Bresnan, J., Cueni, A., Nikitina, T. and Baayen, R. Harald (2007). Predicting the dative alternation. In G. Bouma, I. Krämer and J. Zwarts (Eds), *Cognitive Foundations of Interpretation*. Amsterdam: Royal Netherlands Academy of Arts and Sciences, pp. 69–94.

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

Bybee, J. L. and Scheibman, J. (1999). The effect of usage on degrees of constituency: The reduction of *don't* in English. *Linguistics,* 37(4), 575–96.

Caldwell-Harris, C., Berant, J. and Edelman, S. (2012). Measuring mental entrenchment of phrases with perceptual identification, familiarity ratings, and corpus frequency statistics. In D. S. Divjak and S. Th. Gries (Eds), *Frequency Effects in Language Representation*. Berlin and New York: Mouton de Gruyter, pp. 165–94.

Chomsky, N. A. (1959). A review of B.F. Skinner's Verbal Behavior. *Language,* 35(1), 26–58.

Dąbrowska, E. and Lieven, Elena V. M. (2005). Towards a lexically specific grammar of children's question constructions. *Cognitive Linguistics,* 16(3), 437–74.

Desagulier, G. (2012). Visualizing distances in a set of near-synonyms. In D. Glynn and J. Robinson (Eds), *Polysemy and Synonymy: Corpus Methods and Applications in Cognitive Linguistics*. Amsterdam and Philadelphia: John Benjamins.

Divjak, D. S. (2006). Ways of intending: Delineating and structuring near synonyms. In S. Th. Gries and A. Stefanowitsch (Eds), *Corpora in Cognitive Linguistics: Corpus-based Approaches to Syntax and Lexis*. Berlin and New York: Mouton de Gruyter, pp. 19–56.

Divjak, D. S. and Arppe, A. (2010). Extracting prototypes from corpus data: A distributional account of representing near-synonymous verbs. Paper presented at the interdisciplinary workshop on verbs 'The identification and representation of verb features', Pisa.

Divjak, D. S. and Gries, S. Th. (2006). Ways of trying in Russian: Clustering behavioral profiles. *Corpus Linguistics and Linguistic Theory,* 2(1), 23–60.

— (2008). Clusters in the mind? Converging evidence from near synonymy in Russian. *The Mental Lexicon,* 3(2), 188–213.

— (2009). Corpus-based cognitive semantics: A contrastive study of phasal verbs in English and Russian. In K. Dziwirek and B. Lewandowska-Tomaszczyk (Eds), *Studies in Cognitive Corpus Linguistics*. Frankfurt am Main: Peter Lang, pp. 273–96.

Ellis, N. C. (2007). Language acquisition as rational contingency learning. *Applied Linguistics,* 27(1), 1–24.

— (2012). What can we count in language, and what counts in language acquisition, cognition, and use? In S. Th. Gries and D. S. Divjak (Eds), *Frequency Effects in Language Learning and Processing*. Berlin and New York: Mouton de Gruyter, pp. 7–33.

Ellis, N. C. and Ferreira-Junior, F. (2009). Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics,* 7, 187–220.

Gibbs, R. W. and Matlock, T. (2001). Psycholinguistic perspectives on polysemy. In H. Cuyckens and B. Zawada (Eds), *Polyemy in Cognitive Linguistics*. Amsterdam and Philadelphia: John Benjamins, pp. 213–39.

Glynn, D. (2010). Testing the hypothesis: Objectivity and verification in usage-based cognitive semantics. In D. Glynn and K. Fischer (Eds), *Quantitative Methods in Cognitive Semantics: Corpus-driven Approaches*. Berlin and New York: Mouton de Gruyter, pp. 239–629.

— (2012). The many uses of *run*: Corpus methods and socio-cognitive semantics. In D. Glynn and J. Robinson (Eds), *Polysemy and Synonymy: Corpus Methods and Applications in Cognitive Linguistics*. Amsterdam and Philadelphia: John Benjamins.

Glynn, D. and Fischer, K. (Eds) (2010). *Quantitative Methods in Cognitive Semantics: Corpus-driven Approaches*. Berlin and New York: Mouton de Gruyter.

Glynn, D. and Robinson, J. (Eds) (2012). *Polysemy and Synonymy: Corpus Methods and Applications in Cognitive Linguistics*. Amsterdam and Philadelphia: John Benjamins.

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

— (2011). Corpus evidence of the viability of statistical preemption. *Cognitive Linguistics,* 22(1), 131–53.

Gries, S. Th. (2000). *Multifactorial Analysis in Corpus Linguistics: The Case of Particle Placement*. PhD dissertation, University of Hamburg.

— (2003a). *Multifactorial Analysis in Corpus Linguistics: A Study of Particle Placement*. London and New York: Continuum Press.

— (2003b). Towards a corpus-based identification of prototypical instances of constructions. *Annual Review of Cognitive Linguistics,* 1, 1–27.

— (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics,* 13(4), 403–37.

— (2010a). Dispersions and adjusted frequencies in corpora: Further explorations. In S. Th. Gries, S. Wulff and M. Davies (Eds), *Corpus Linguistic Applications: Current Studies, New Directions*. Amsterdam: Rodopi, pp. 197–212.

— (2010b). Behavioral profiles: A fine-grained and quantitative approach in corpus-based lexical semantics. *The Mental Lexicon,* 5(3), 323–46.

— (2010c). Corpus linguistics and theoretical linguistics: A love-hate relationship? Not necessarily . . . *International Journal of Corpus Linguistics,* 15(3), 327–43.

—(2012). Frequencies, probabilities, association measures in usage-/exemplar-based linguistics: Some necessary clarifications. *Studies in Language,* 36(3), 477–510.
— (2013a). *Statistics for Linguistics with R*. (2nd rev. and exp. ed.). Berlin and New York: Mouton de Gruyter.
— (2013b). 50-something years of work on collocations: What is or should be next. . . . *International Journal of Corpus Linguistics,* 18(1), 137–65.
Gries, S. Th. and Stefanowitsch, A. (2004a). Extending collostructional analysis: A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics,* 9(1), 97–129.
— (2004b). Co-varying collexemes in the *into*-causative. In M. Achard and S. Kemmer (Eds), *Language, Culture, and Mind*. Stanford, CA: CSLI, pp. 225–36.
— (2010). Cluster analysis and the identification of collexeme classes. In S. Rice and J. Newman (Eds), *Empirical and Experimental Methods in Cognitive/functional Research*. Stanford, CA: CSLI, pp. 73–90.
Gries, S. Th. and Wulff, S. (2005). Do foreign language learners also have constructions? Evidence from priming, sorting, and corpora. *Annual Review of Cognitive Linguistics,* 3, 182–200.
— (2009). Psycholinguistic and corpus linguistic evidence for L2 constructions. *Annual Review of Cognitive Linguistics,* 7, 163–86.
— (in progress). *Corpora in Cognitive Linguistics*. Amsterdam and Philadelphia: John Benjamins.
Gries, S. Th., Hampe, B. and Schönefeld, D. (2005). Converging evidence: Bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics,* 16(4), 635–76.
— (2010). Converging evidence II: More on the association of verbs and constructions. In S. Rice and J. Newman (Eds), *Empirical and Experimental Methods in Cognitive/functional Research*. Stanford, CA: CSLI, pp. 59–72.
Hay, J. B. R. Baayen, H. (2003). Phonotactics, parsing and productivity. *Italian Journal of Linguistics,* 1, 99–130.
Hilpert, M. (2006). Distinctive collexeme analysis and diachrony. *Corpus Linguistics and Linguistic Theory,* 2(2), 243–57.
— (2008). *Germanic Future Constructions: A Usage-based Approach to Language Change*. Amsterdam and Philadelphia: John Benjamins.
Janda, L. A. (Ed.) (2013). *Cognitive Linguistics: The Quantitative Turn*. Berlin and New York: Mouyton de Gruyter.
Janda, L. A. and Solovyev, S. D. (2009). What constructional profiles reveal about synonymy: A case study of Russian words for SADNESS and HAPPINESS. *Cognitive Linguistics,* 20(2), 367–93.
Joseph, B. D. (2004). On change in language and change in language. *Language,* 80(3), 381–3.
— (2008). Last scene of all . . . *Language,* 84(4), 686–90.
Kisnher, J. M. and Raymond, W. Gibbs Jr (1996). How *just* gets its meanings: Polysemy and context in psychological semantics. *Language and Speech,* 39(1), 19–36.
Lakoff, G. (1990). The invariance hypothesis: Is abstract reason based on image schemas? *Cognitive Linguistics,* 1(1), 39–74.
Langacker, R. W. (1987). *Foundations of Cognitive Grammar, Vol. 1: Theoretical Prerequisites*. Stanford, CA: Stanford University Press.
Lemke, S., Tremblay, A. and Tucker, Benjamin V. (2009). Function words of lexical bundles: The relation of frequency and reduction. *Proceedings of Meetings on Acoustics,* 6, 060009.
Levshina, N. (in prep). Lexical fields and constructional spaces: A quantitative corpus-based model of semantics.

Levshina, N., Geeraerts, D. and Speelman, D. (2012). Dutch causative constructions with *doen* and *laten*: Quantification of meaning and meaning of quantification. In D. Glynn and J. Robinson (Eds), *Polysemy and Synonymy: Corpus Methods and Applications in Cognitive Linguistics*. Amsterdam and Philadelphia: John Benjamins.

— (2013). Mapping constructional spaces: A contrastive analysis of English and Dutch analytic causatives. *Linguistics,* 51(4), 825–54.

Mukherjee, J. (2007). Corpus linguistics and linguistic theory: General nouns and general issues. *International Journal of Corpus Linguistics,* 12(1), 131–47.

Raymond, W. D. and Brown, E. L. (2012). Are effects of word frequency effects of context of use? An analysis of initial fricative reduction in Spanish. In S. Th. Gries and D. S. Divjak (Eds), *Frequency Effects in Language Learning and Processing*. Berlin and New York: Mouton de Gruyter, pp. 35–52.

Rice, S. and Newman, J. (Eds) (2010). *Empirical and Experimental Methods in Cognitive/functional Research*. Stanford, CA: CSLI.

Sandra, D. and Rice, S. (1995). Network analyses of prepositional meaning: Mirroring whose mind – the linguist's or the language user's? *Cognitive Linguistics,* 6(1), 89–130.

Schmid, H.-J. (1993). Cottage *and* co., idea, start *vs.* begin. Tübingen: Max Niemeyer.

Schönefeld, D. (Ed.) (2011). *Converging Evidence: Methodological and Theoretical Issues for Linguistic Research.* Amsterdam and Philadelphia: John Benjamins.

Snider, N. and Arnon, I. (2012). A unified lexicon and grammar? Compositional and non-compositional phrases in the lexicon. In D. S. Divjak and S. Th. Gries (Eds), *Frequency Effects in Language Representation*. Berlin and New York: Mouton de Gruyter, pp. 127–63.

Stefanowitsch, A. (2010a). Cognitive linguistics meets the corpus. In M. Brdar, S. Th. Gries and M. Žic Fuchs (Eds), *Cognitive Linguistics: Convergence and Expansion*. Amsterdam and Philadelphia: John Benjamins, pp. 257–90.

— (2010b). Empirical cognitive semantics: Some thoughts. In D. Glynn and K. Fischer (Eds), *Quantitative Methods in Cognitive Semantics: Corpus-driven Approaches*. Berlin and New York: Mouton de Gruyter, pp. 355–80.

— (2011). Constructional preemption by contextual mismatch: A corpus-linguistic investigation. *Cognitive Linguistics,* 22(1), 107–29.

Stefanowitsch, A. and Gries, S. Th. (2003). Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics,* 8(2), 209–43.

— (2005). Covarying collexemes. *Corpus Linguistics and Linguistic Theory,* 1(1), 1–43.

Stefanowitsch, A. and Gries, S. Th. (Eds) (2006). *Corpus-based Approaches to Metaphor and Metonymy*. Berlin and New York: Mouton de Gruyter.

Street, J. A. and Dąbrowska, E. (2010). More individual differences in language attainment: How much do adult native speakers of English know about passives and quantifiers? *Lingua,* 120(8), 2080–94.

Szmrecsanyi, B. (2006). *Morphosyntactic persistence in spoken English: [. . .]*. Berlin and New York: Mouton de Gruyter.

Talmy, L. (2000). *Toward a Cognitive Semantics. Vol. 1. Concept Structuring Systems. Vol. 2. Typology and Process in Concept Structuring*. Cambridge, MA: MIT Press.

Teddiman, L. (2012). Conversion and the lexicon: Comparing evidence from corpora and experimentation. In D. S. Divjak and S. Th. Gries (Eds), *Frequency Effects in Language Representation*. Berlin and New York: Mouton de Gruyter, pp. 235–54.

Theijssen, D., Bosch, Louis ten, Boves, L., Cranen, B. and Halteren, Hans van (to appear). Choosing alternatives: Using Bayesian networks and memory-based learning to study the dative alternation. *Corpus Linguistics and Linguistic Theory*.

Thompson, S. A. and Koide, Y. (1987). Iconicity and 'indirect objects' in English. *Journal of Pragmatics*, 11(3), 309–406.
Wulff, S. (2008). *Rethinking Idiomaticity: A Usage-based Approach*. London and New York: Continuum.
Zeldes, A. (2012). *Productivity in Argument Selection: A Usage-based Approach to Lexical Choice in Syntactic Slots*. PhD dissertation, Humboldt University Berlin.
Zeschel, A. (2012). *Incipient Productivity: A Construction-based Approach to Linguistic Creativity*. Berlin and New York: Mouton de Gruyter.