36

Corpus Approaches

Stefan Th. Gries

36.1 Introduction

36.1.1 General 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) cognitive 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; and analogy and similarity.

Even though these notions all involve human cognition and have been addressed with quite some empirical rigor in psychology or psycholinguistics, much early cognitive linguistic research was based on introspection just as much as the generative approach against which much of it was arguing. However, in the last twenty to twenty-five years or so, there has been a greater recognition of the problems that arise when linguists do not (try to) back up their often far-reaching claims and theories with robust data. With regard to polysemy networks, for instance, Sandra and Rice (1995) have been a wake-up call in terms of 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 or Gries and Ellis 2015 for recent overviews). This chapter provides an overview of how corpus data and statistical methods are used in increasingly sophisticated ways in cognitive linguistics. In section 36.1.2, I briefly introduce some of the most central corpus-linguistic notions as well as how methods and goals of corpus-linguistic research are related to cognitive linguistics before turning to examples and applications in sections 36.2 to 36.4. Specifically and for expository reasons alone, I discuss (more) lexical examples in section 36.2 and (more) syntactic examples in section 36.3; I will then turn to selected applications of quantitative corpus linguistics in other fields in section 36.4. Section 36.5 mentions some necessary future developments.

36.1.2 Corpus linguistics: Methods and Goals

Corpus linguistics is the study of data in corpora. The notion of a corpus can be considered a prototype category with the prototype being a collection of machine-readable files that contain text and/or transcribed speech that were produced in a natural communicative setting and that are supposed to be representative of a certain language, dialect, variety, etc.; often, corpus files are stored in Unicode encodings (so that all sorts of different orthographies can be appropriately represented) and come with some form of markup (e.g. information about the source of the text) as well as annotation (e.g. linguistic information such as part-of-speech tagging, lemmatization, etc. added to the text, often in the form of XML annotation). However, there are many corpora that differ from the above prototype along one or more dimensions. The most important of these dimensions are probably *size* (from a few narratives narrated in an under-documented or even already extinct language to corpora of many billions of words) and *degree of naturalness of the data* they contain (from completely spontaneous dialog between two speakers to experimental, highly constrained situations).

From one point of view, the technical aspects of corpus-linguistic analyses are quite simple and limited: nearly all kinds of corpus analysis are based on using specialized corpus software or, more usefully, programming languages, to retrieve from corpora examples of words or, more generally, constructions (in the Construction Grammar sense of the term), which are then analyzed in one of two different ways:

1. Analyses in which the construction(s) of interest is/are analyzed in a fairly or completely decontextualized manner. That is, one might consider *frequency lists*: how often do words *x, y, z* occur in (parts of) a corpus? *Dispersion*: how evenly are syntactic constructions *a, b, c* distributed in a corpus (e.g. in the paternal input to a child)? *Collocation/colligation*: which words occur how often in, or with, another construction? And so on. In many such cases, there is no more detailed analysis of the (contexts of the) uses of the construction of interest – rather, the

analysis of frequencies is based on observed frequencies or (conditional) probabilities of occurrence, or specialized measures of dispersion or co-occurrence/association.

2. Analyses in which the construction(s) of interest is/are studied by means of context, usually on the basis of *concordances*, that is, displays of instances of (a) construction(s) and the detailed annotation of contextual and/or (b) linguistic (phonological, morphological, syntactic, semantic) features.

On the one hand, this might not only seem like quite a limited number of methods but also like a limited number of methods providing nothing but frequencies of occurrence or co-occurrence of character strings. In a sense, that is correct. On the other hand, however, the above corpus methods actually provide an immensely wide range of useful applications for cognitive and/or usage-based linguistics because of what may be the most important assumption uniting usage-based and corpus linguistics, the so-called *distributional hypothesis*. This assumption/working hypothesis assumes that the similarity of linguistic elements with regard to their functional characteristics – semantic, pragmatic, etc. – is reflected in their distributional similarity/characteristics. Harris (1970: 785f.) put it best:

> If we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference of meaning correlates with difference of distribution.

This means that most of the central notions of cognitive/usage-based linguistics have close analogs in corpus linguistics or can be operationalized (more or less directly) on the basis of corpus data. Consider, for instance the following notional parallels between cognitive and corpus linguistics: just as cognitive linguists began exploring the notion that there may not be a clear-cut divide between lexis and syntax (e.g. Langacker 1987), corpus linguists independently began to do the same (e.g. Sinclair 1991). In the same vein, just as Langacker (1987: 42, 57) rejected the rule-list fallacy and proposed the notion of automatically deployable units, corpus linguists were discussing Sinclair's Idiom Principle, according to which speakers make use of semi-preconstructed expressions. Just as cognitive linguists were adopting the notion of (argument structure) constructions (Goldberg 1995, 2006: 5), corpus linguists were considering Hunston and Francis's (2000: 37) patterns, etc.

The mutually beneficial parallels do not end with the above more theoretical correspondences. As usage-based linguistics is becoming more empirically rigorous and psycholinguistic, it finally takes more notice of psychological and psycholinguistic findings and relies more on exactly the kinds of data that corpora provide: token and type frequencies of (co-)occurrence. For instance, learning and cognition – linguistic and otherwise – are massively affected by frequency, recency, and context of usage (e.g. Ebbinghaus 1885, Bartlett 1932 [1967], Anderson 2000), which means that "Learners FIGURE language out: their task is, in essence, to learn the probability distribution P (interpretation|cue, context), the probability of an interpretation given a formal cue in a particular context, a mapping from form to meaning conditioned by context" (Ellis 2006b: 8). Paraphrasing Ellis immediately leads to two recognitions. First, a psychologically/cognitively informed approach to language presupposes and requires at the same time the precept that meaning is ultimately distributional in nature, and that is exactly what corpus frequencies provide proof of, including the insights that:

1. High(er) observed *token frequency of occurrence* (e.g. how often does a construction occur?) or *conditional probabilities* (e.g. how often does a construction occur given the presence of another construction?) are correlated with entrenchment (via the power law of learning), predictability, phonetic reduction, early acquisition, reaction times (in, say, lexical decision tasks, word and picture naming tasks) etc.
2. High(er) *type frequencies or frequencies of co-occurrence* (e.g. how many different word types occur in particular slots of constructions?) are essentially an operationalization of productivity, and thereby are correlated with generalization in acquisition, grammaticalization effects, higher predictability or lower degrees of surprisal, etc.
3. *High frequencies of co-occurrence* or *high degrees of association* are often associated with high degrees of contingency (often of form and function as in the Competition Model; see MacWhinney 1987a) and again higher predictability/lower surprisal.

Second, the reverse perspective works just as well: rather than going from explaining how different kinds of corpus data are related to central notions from usage-based linguistics (as above), one can show how many (other) cognitive linguistic notions are readily operationalized in corpus terms, including the insights that:

1. *Prototypicality* and *basic-level category status* are often correlated with token frequency of occurrence, diversity of contexts, and high diachronic and or acquisitional frequency as measured in corpora.
2. *Recency* of usage events can be operationalized on the basis of dispersion of items in corpus data: how regular, or even-spaced, are occurrences of a particular expression, or how rare but then clumpy are they in a corpus?
3. Salience of (expressions in) usage events can be operationalized on the basis of how predictable (an expression in) a usage event is, for instance on the basis of conditional probabilities or of something derived from them.

4. Context and form-function contingency are available from concordances (generated from, ideally, highly annotated corpora) and subsequent association measures that quantify the strength and, ideally, the direction of association of elements.
5. More generally, *learning* is not only based on recognizing contingencies from the masses of exemplars of usage events stored, but also driven by prediction errors: we notice, and learn from, more when our predictions/expectations (about what linguistic element comes next, what context a linguistic element is used in etc.) are incorrect than when they are confirmed, which is something that can be operationalized as (negative binary logs of) conditional probabilities of elements in corpus data, that is, surprisal, etc.

In sum, while corpus linguistics 'only' provides distributional data, such distributional data are exactly what is relevant to the cognitive-/usage-based linguist via (i) the distributional hypothesis in general and via (ii) the close correspondences between, and often very direct operationalizations of, cognitive notions into corpus-based distributional patterns. In the following sections, I will discuss a variety of cognitive linguistic applications that involve (various combinations of) the above-mentioned kinds corpus linguistic data and their statistical exploration.

36.2 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 and collocations 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 often 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 with *the* a lot). Early examples of the study of co-occurrence in cognitive linguistics are Schmid (1993), Kishner and Gibbs (1996) on *just*, and Gibbs and Matlock (2001) on *make*. While these studies of collocation and colligation were groundbreaking at the time, they were still largely monofactorial in nature: Uses of (senses of) words were annotated for, and cross-tabulated with, co-occurrence patterns, but often no deeper quantitative analyses was conducted; a few selected examples will be discussed in section 36.2.1. The current state of the art, however, is that multidimensional co-occurrence data are gathered and statistically analyzed accordingly. Gries (2010b) distinguishes two different ways in which analyses can be multidimensional, which will be exemplified in sections 36.2.2 and 36.2.3.

36.2.1 Early Univariate Studies

Much early corpus work in cognitive linguistics was concerned with conceptual metaphor and metonymy and was based on concordance data resulting from (i) completely manual searches, (ii) searches for source and/or target domain expressions, and/or (iii) searches in corpora annotated for conceptual mappings; such searches may be based on individual expressions or sets of expressions (either defined a priori, on the basis of lists, or using some exploratory procedure). Stefanowitsch's (2006b) metaphorical pattern analysis is an example: for several basic emotions, he chose a lexeme and retrieved up to 1000 matches from the British National Corpus, uncovering altogether 1443 metaphorical patterns, which are then classified in terms of the (kinds of) metaphors they instantiate and compared to a representative non-corpus-based study. The results indicate that this kind of analysis is superior to the mere introspective/opportunistic listing of metaphors, in particular in how this kind of analysis can be used to discover many metaphors that more traditional study did not find (see Stefanowitsch 2004 for a similar analysis contrasting Happiness metaphors in English and German and Hilpert 2006a for a similar study of metonymies with *eye*). Finally, Gries (2006) is a corpus-based study of the many metaphorically and metonymically related senses of 815 instances of the highly polysemous verb *to run* in British and American English.

Studies of the above kind helped pave the way to quantitatively more complex analyses exploring the distribution of linguistic items in many different dimensions.

36.2.2 Multidimensional$_1$ Approaches: Behavioral 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, contextual ones – and all of these dimensions are used in a statistical analysis at the same time, but *separately*. One example for this approach that has become more widely used is the behavioral profile (BP) approach (see Gries 2010b for an overview). Concordance lines are annotated for many features, 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 them. Gries (2006) applied this method to cluster senses of *to run*, and 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 (i.e. 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 behavioral 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 analyzed the similarity of BP co-occurrence percentages with cluster analyses and some follow-up exploration and 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 sets of sorting and gap-filling experiments with native speakers of Russian, lending further support to the method in particular, but also to the idea of converging evidence in general. Since then BPs have been used in more specialized settings, as when Janda and Solovyev (2009) use a downsized version of BP data – the constructional profile, the relative frequency distribution of the grammatical constructions a word occurs in – to explore synonyms, but also in more general settings, as when Divjak and Gries (2009) compare the use of phasal verbs in English (*begin* versus *start*) and Russian (*načinat'/načat', načinat'sja/načat'sja*, and *stat'*). Among other things, they find that English speakers' choices are driven by semantic characteristics of the beginners and beginnees whereas Russian speakers' choices are more driven by aspectual and argument-structural characteristics.

36.2.3 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 characteristics – 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, etc. – behaves in each of many dimensions such as what are the percentages with which sense x has different kinds of subjects or what are the percentages with which sense x has different kinds of objects? 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. However, that does not include 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 many combinations of features are rare or even unattested.

Two types of multidimensional$_2$ applications are particularly interesting. First, there are 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 (2010b) study of *bother*. Glynn followed the work discussed in section 36.2.2 and annotated uses of *bother* for a large number of features and applied MCAs to different parts of the multidimensional frequency table to find different clusters and "semantically motivated distinction[s] between two sets of syntactic patterns" (Glynn 2010b: 256) – an agentive and a predicative construction. In order to test the patterns suggested by the exploratory tool, Glynn then added a second type of multidimensional$_2$ application, confirmatory approaches based on regression analyses, 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. Similar applications in the domain of semantics include Glynn (2014a) and Desagulier (2014).

Additional examples for similar multidimensional$_2$ applications involve binary as well as multinomial or polytomous logistic regressions. As for the former, Gries and Deshors (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 allow best 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 Gries (2003b), who uses linear discriminant analysis to the same end.

Regardless of which multidimensional approach is chosen, the combination of comprehensive annotation and multifactorial/-variate analysis has proven to yield insightful results regarding a variety of the abovementioned 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 versus 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, cf. Glynn and Fischer (2010) and Glynn and Robinson (2014).

36.3 Syntax-lexis, with an Emphasis on Syntax

Somewhat 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 the use of regular expressions to retrieve (parts of) words, part of speech tags, 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.

36.3.1 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, Gries and Stefanowitsch 2004a, 2004b, Stefanowitsch and Gries 2005). 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 and 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 (see section 36.4.2), constructional change over time (Hilpert 2006b, 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 et al. (2005) demonstrated how collexeme analysis outperforms frequency and conditional probabilities as predictors of subjects' behavior 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, 2015a, 2015b) for comprehensive discussion and rebuttals of recent critiques of the method. Lastly, collostructions have been coupled with more advanced statistical tools – such as cluster analysis or correspondence analysis – to discover subsenses of constructions (cf. Gries and Stefanowitsch 2010), or they have been refined better to incorporate senses of constructions (e.g. Perek 2014) or senses of verbs (e.g. Bernolet and Colleman, in prep.).

36.3.2 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 36.2, 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.

Among the very 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*. He annotated examples of both constructions from the British National Corpus (BNC) for a large number of phonological, morphological, syntactic, semantic, and discourse-functional parameters and used a classifier – linear discriminant analysis – to identify the factors that make speakers choose one construction over another in a particular discourse context, to discuss their implications for language production, and to identify prototypical instances of both constructions. Since then, this type of approach – multifactorial modeling 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) and Shank et al. (2016) are two cases in point using logistic regressions (the latter being a diachronic study) (see Levshina, Geeraerts, and Speelman 2014 for the additional tool of classification and regression trees. Gries

(2003b) showed that the predictions of such methods correlate very strongly with results from acceptability ratings.

Exploratory approaches in this domain 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 (approximately 10 m words) for English and an equally large sample from the Twente and the Leuven News corpora, Levshina, Geeraerts, and Speelman (2013) retrieved approximately 4000 examples of causatives in 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 in the follow-up there is discussion about how both languages' data points are located differently in causative space.

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

As mentioned above and for purely expository reasons, sections 36.2 and 36.3 in this chapter upheld a distinction that many cognitive and corpus linguists do not 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 the concept of construction that an expression being 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 of multiword units (MWUs, also often called *lexical bundles*), that is, expression 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 of point, become retained in speakers' minds and, thus, most likely also give 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 et al. (2012) based on the comparison of transitional probabilities in corpus data and eye-tracking data (cf., for more discussion, Snider and Arnon 2012 or 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, and so on, all these are areas where state-of-the-art quantitative corpus linguistics can be very useful (for more examples, cf. Stefanowitsch 2010 and the papers in Gries and Stefanowitsch 2006, Rice and Newman 2010, Schönefeld 2011, Divjak and Gries 2012, and Gries and Divjak 2012).

36.4 Other Linguistic Subdisciplines

36.4.1 Structural Subdisciplines

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 considerabe body of corpus-based cognitive linguistic research in domains other than ‘pure’ syntax or lexis. While space does not permit an exhaustive discussion, the following highlights some examples.

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 of 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.

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 versus non-decompositional approaches (cf. Hay and Baayen 2003 or Antić 2012 for a more recent contribution).

The following summarizes a few other studies that involve, or are defined by, morphological elements. Berez and Gries (2010) explored the factors that trigger the ab-/presence of the middle marker *d* in iterative verbs on the basis of a small corpus of Dena'ina narratives. 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.

Teddiman (2012) showed how subjects' decision 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. For instance and on the whole, words such as *pipe* and *drive* (mostly used nominally and verbally respectively) were typically assigned to be nouns and verbs respectively.

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 of 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 (where recognizability was operationalized in a corpus-based way). Again, this corpus-informed work sheds light on a phenomenon that traditional morphology finds difficult to cope with.

Finally, Sokolova et al. (2012) as well as Backus and Mos (2011) connect morphology and syntax. Using their variant of BPs, constructional profiles, the former explore nearly 2000 examples of the Russian locative alternation with грузить and three of its prefixed forms from the Modern subcorpus of the Russian National Corpus. They model the constructional choice using three predictors – prefix, number of participants, finite/participle form of the verb – in a logistic regression and find, among other things, significant differences between the four different verbs, which is particularly interesting since the "three perfectives are traditionally considered to bear semantically 'empty' prefixes" (2011: 67); thus, the corpus-based approach goes against received wisdom and shows that the meanings of the verbs and the constructions interact. The latter study is concerned with the productivity and similarity of two Dutch potentiality constructions – a derivational morpheme (*-baar*) and a copula construction (SUBJ $\text{COP}_{\text{finite}}$ *te* INF) and is a nice example of how corpus data are used complementarily with other kinds of data, here acceptability judgments. They report the results of a distinctive collexeme analysis to determine which verbs prefer which of the two constructions in the Corpus of Spoken Dutch and follow this result up with a judgment experiment to probe more deeply into seemingly productive uses of the constructions. They found converging evidence such that acceptability is often correlated with corpus frequencies and preferences (see the chapters in Schönefeld 2011 for more examples of converging evidence).

36.4.2 Other Subdisciplines

Apart from the structurally motivated subdisciplines of phonology, morphology, lexis, and syntax, corpus-based work in cognitive/usage-based linguistics has also had a particular impact on the following three areas, each of which is influenced particularly by one central figure/research group and which will be discussed very briefly in what follows: first language acquisition, second/foreign language acquisition, and cognitive sociolinguistics.

In L1 acquisition research, the work done by Tomasello, Lieven, and colleagues has been among the most influential corpus-based work in cognitive linguistics (see Tomasello 2010 for a fairly recent overview). One currently ‘hot’ area is concerned with how children learn what not to say, that is, how they learn to avoid overgeneralizations – by negative entrenchment or statistical preemption (see Stefanowitsch 2011, Goldberg 2011, Ambridge et al. 2012, Robenalt and Goldberg 2015, among others). In addition, this field is also slowly embracing more computational methods, such as Dąbrowska and Lieven (2005) on the development of early syntax using the traceback method, or larger data bases, such as Behrens (2006), who explores parts-of-speech information as well as 300 K NPs and 200 K VPs with regard to how distributions in children’s data over time come to approximate the (stable) distributions in adult data.

In L2-acquisition/foreign-language learning research, the most influential work is by Nick Ellis and colleagues. One early influential study is Ellis and Ferreira-Junior (2009), who retrieve all instances of three argument structure constructions from the ESL corpus of the European Science Foundation project to compute type frequencies, type-token distributions, and collexeme strengths (of verbs and constructions) to test for Zipfian distributions and identify first-learned and path-breaking verbs. Gries and Wulff (2005, 2009) are studies of alternations (the dative alternation and *to*/*-ing* complementation) that combine corpus data with experimental results; Ellis, O’Donnell, and Römer (2014b) correlate results from a gap-filling experiment with German, Czech, and Spanish learners of English to the frequencies of verbs in the same constructions (entrenchment), the associations of these verbs to the constructions (contingency), and their semantic prototypicality using multiple regression. They find that each factor makes its own significant contribution to the frequency with which learners provide verbs for constructions.

Last but not least, there is now some interest in cognitive sociolinguistics, mostly stimulated by work done by Geeraerts and colleagues. Studies such as Glynn (2014b) or Levshina, Geeraerts, and Speelman (2014) argue in favor of adding predictors covering dialectal, geographic, thematic variability to their statistical analyses. While the results reported in such studies indicate that including these dimensions in statistical modeling increases the overall amount of explained variability in the data, it seems to me as if it still needs to be shown to what degree such findings also inform the cognitive aspects of the phenomena thus studied; Pütz, Robinson, and Reif (2014) is a recent interesting collection of work representative of this approach.

36.5 Concluding Remarks and Future Developments

As the previous sections have demonstrated, corpus linguistic methods and subsequent statistical analysis have become very important for cognitive and exemplar-/usage-based linguistics. The type of exemplar-based approaches that many cognitive linguists now embrace are particularly compatible with the distributional data that corpora provide, and cognitive and corpus linguists have independently arrived at many shared notions and perspectives. In this final section, I would like very briefly to provide some comments on where I think cognitive linguistics can and should evolve and mature further by incorporating insights from quantitative corpus linguistics.

36.5.1 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 bidirectional 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 (2006a) 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 2013b, 2015a), 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, 2010a) discussed ways to measure this as well as first results that indicate that, sometimes, dispersion is a better predictor of experimental results than frequency.

The field should also consider further individual differences, which will require that researchers take seriously how corpora represent speakers' individual contributions to the data. Studies such as Street and Dąbrowska (2010) or Caldwell-Harris, Berant, and Edelman (2012) and others show clearly that the 'native speaker,' about which all linguistic theories like to generalize, is merely a convenient fiction, given the huge individual diversity that both corpus and experimental data reveal very clearly.

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

36.5.2 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 multilevel) modeling is becoming more frequent, since it allows the analyst to handle subject/speaker-specific (see above) and, for example, word-specific variation, as well as unbalanced data, much better than traditional regression tools; in addition, other methods such as multimodel inferencing and random forests promise to address potential shortcoming of often very noisy and collinear datasets (see Kuperman and Bresnan 2012 or Gries 2015b for the former as well as Matsuki, Kuperman, and Van Dyke 2016 and Deshors and Gries, in prep., for the latter. On the other hand, new classification tools such as Bayesian network and memory-based learning (cf. Theijssen et al. 2013), with its capacity for the modeling of causal effects, in a way reminiscent of structural equation modeling, and naïve discriminative learning (cf. Baayen 2010), with its higher degree of cognitive realism, are becoming important promising new alternatives; in addition, the modeling of nonlinear relations between predictors and responses by means of, say, generalized additive models, is slowly becoming more frequent in linguistics (see, e.g., Wieling, Nerbonne, and Baayen 2011). Hopefully it will also catch on in cognitive linguistics. Finally, I hope that exploratory/bottom-up techniques will become more frequently used. While cluster and correspondence analyses are already in more widespread use, methods such as network analysis (see Ellis et al. 2014a for a very interesting application of graph-based algorithms to verbs in slots of constructions) or longitudinal connectionist or exemplar-based simulations hold much promise for cognitively more realistic statistical evaluations; for an interesting example of a longitudinal corpus study, on German *was… für* (‘what kind of…’ questions, see Steinkrauss 2011).

While this chapter could only provide the briefest of overviews of the impact that corpora and quantitative methods have had on cognitive linguistics, it is probably fair to say that such methods are taking the field by storm. It is to be hoped that this development and maturation of the field continues as individual scholars increase their repertoire of corpus and quantitative skills and their engagement with experimental data, and as more and more fruitful connections with neighboring disciplines – e.g. corpus linguistics or psycholinguistics – provide opportunities for interdisciplinary research.