Is quantitative empirical research possible for the study of semantics? More specifically, can we use corpus data to produce testable and falsifiable results for the description of meaning? If so, what are the advantages and what are the limitations of such an approach? Such questions are often heard and it would seem they are behind much of the reticence and scepticism for corpus-driven techniques in the broader Cognitive Linguistic community. This volume grew out of the workshop Usage-Based and Quantitative Methodology in Cognitive Semantics at the International Cognitive Linguistics Conference 2007, in Kraków, where the participants were asked to answer these very questions. This book both critiques and supports the application of such techniques to the study of meaning in language.

1. Cognitive Linguistics – A necessarily empirical approach

Cognitive Linguistics is beginning to realise the implications of its own theoretical framework. When Langacker (1987, 1988) first outlined the theory of entrenchment and his usage-based model, he may not have appreciated the extent to which these proposals would be implied in the trends of 21st century linguistics. Indeed, Langacker’s theoretical proposal arguably set the stage for a major paradigm shift in linguistics, from theory-driven to empirical research. Similarly, when Lakoff (1982, 1987) argued that we need to approach language holistically, that categorisation based on human experience is the foundational structure of meaning, he is likely to have underestimated the implications of this move. A holistic approach to meaning goes far beyond adding pragmatics to semantic analysis – it entails that everything speakers know of the world, including the whole domain of sociolinguistics, is necessarily drawn into the mainstream of language description.

It is due to these theoretical tenets that the roots of Cognitive Linguistics were, and inevitably so, an empirical movement seeking to account for the
totality of language as a socio-cognitive phenomenon. The non-empirical emphasis of earlier research can be attributed to methodological practices inherited from the Structuralist and Generative (or Mentalist) schools. These schools assume an object of study based on society’s *langue* or the ideal speaker’s competence, and as such, their methods can rely exclusively on introspection.

Cognitive Linguistics’ model of language necessarily entails empirical methodology. In fact, Fillmore (1985), Talmy (1985), Lakoff (1987), and Langacker (1987) can be said to have produced one of the first data-oriented models of language. Their model dismantles a hundred years of fervent linguistic theorising. They deconstruct the Structuralist distinction of semantics and pragmatics, demonstrate the arbitrariness of the Modularist division between lexis and syntax, and argue that language structure is a result of usage, based on general cognitive capacities. The object of study is finally language use in all its complexity, rather than *langue* or ‘competence’, lexis or syntax, and semantics or pragmatics.

The model of language proposed by Cognitive Linguistics is so completely simple that it places the emphasis squarely on method and data. Rather than simplifying the object of study by carving off its complexities with hypothetical modules of language structure, it lands the linguist in the midst of a chaotic phenomenon that is the nature of all socially structured systems. The original cognitive theoreticians did this in dialectic with Structuralism and Mentalism. Most linguists today would agree that the Cognitivists succeeded in demonstrating how such reductionist theories failed. But did the Cognitivists offer a viable alternative? In some sense, they took us out of the frying pan and into the fire. Firstly, claiming that grammar is semantics-driven and that linguistic meaning extends beyond traditional semantics to all world knowledge exponentially increases the complexity of the object of study. Secondly, both Structuralism and Mentalism claimed that early Cognitive Linguistics offered no means for testing hypotheses or falsifying analyses. They were right. Lakoff’s (1987) claim that co-occurrence of semantic phenomena is the basis of Cognitive Linguistics’ scientific method is largely vacuous unless results obtained with this method can be falsified. Likewise, it is all very well for Langacker (1987) to argue that grammaticality is relative and varies from person to person. But if this is the case, how can we test proposed grammatical structures?

Despite the fact that neither Lakoff nor Langacker developed an empirical line of research, together their proposals establish the cornerstone of
true empirical linguistic enquiry. Lakoff argued that evidence for his proposals comes from the co-occurrence of linguistic phenomena. Reframed, his argument was for inductive analysis, a method of analysis which is the norm in the social sciences. Making generalisations based on a sample, then extrapolating those generalisations to the population is the basis of inductive scientific research, and the only viable method for the social sciences. Yet, without rigorous and testable techniques for establishing how representative or reliable a generalisation is, the ‘results’ remain merely hypotheses based on very small samples. Indeed, as was typical at the time, a sample was no more than the internal language knowledge of the linguist and, perhaps, a few colleagues.

It is here that Langacker’s theory of entrenchment and the usage-based model step in – they offer an operationalisation of grammar and grammaticality. In empirical research, operationalisation is a basic analytical tool – the definition of an object of study by how one can measure it. Once dealing with a measurable object of study, one may test the accuracy of its description and falsify predictions made about it. The ability to falsify results is crucial if language analysis is to progress without grammaticality tests. Langacker’s theory allows us, therefore, to test the accuracy of generalisations based on Lakoff’s identification of co-occurring linguistic phenomena. These two points are crucial for quantitative corpus-driven semantic research and we return to these points below.

Having established that the theory of Cognitive Linguistics is inherently empirical, can we say, however, that the paradigm is empirical in practice? To some extent, empiricism, be that based upon the experimental data of psycholinguistics or upon the collected data of corpus linguistics, is a new turn for Cognitive Linguistics. Yet, despite the recent expansion of empirical methods, both corpus-driven and experimental research go back to the origins of Cognitive Linguistics.²

Since the very beginning of the movement, a small yet important number of scientists have taken very seriously the methodological implications of the cognitive linguistic model of language. Two early works, by scholars whose names were crucial in the foundational years of the paradigm, are testimony to this. Firstly, Dirven et al. (1982) is a corpus-driven study of depth and maturity that would hold its own today. Secondly, Lehrer (1982) is a linguistic experiment whose design and thoroughness exceed many of our current studies. The former investigation led, directly and indirectly, to the development of corpus Cognitive Linguistics in Belgium and Germany. The latter work, recently republished by Oxford University Press (Lehrer
2009), is enjoying renewed attention a quarter of a century later. Although the experimental track is equally important to the development of Cognitive Linguistics, it lies beyond the purview of this volume. The current volume belongs to the tradition, established with Dirven et al. (1982), focusing on methods employing corpus data.

From Dirven et al. (1982), via Dirven & Taylor (1988), Schulze (1988), Rudzka-Ostyn (1989, 1995, 1996) and Goossens (1990) up until Geeraerts (1993b), Schmid (1993, 1997, 2000) and Geeraerts et al. (1994, 1999), corpus-driven Cognitive Linguistics grew slowly but surely. During the 1990s, the movement was given added weight by the demonstration of the inadequacies of the analytical techniques current at the time. The Structuralist and Mentalist-Modularist assumptions about how language works, assumptions that still drove the analytical techniques of early Cognitive Linguists, were shown to be ineffective. Geeraerts (1993a) demonstrated theoretically that the study of polysemy still made Structuralist assumptions about how meaning is organised. Essentially, the kind of radial network analysis popularised by Lakoff (1987) sought to identify discrete lexical senses where there was no reason to suppose that such discrete senses exist:

The tremendous flexibility that we observe in lexical semantics suggests a procedural (or perhaps ‘processual’) rather than a reified conception of meaning; instead of meanings as things, meaning as a process of sense creation would seem to become our primary focus of attention. Geeraerts (1993a: 260)

Two years later, Sandra & Rice’s (1995) experimental study ‘Mirroring whose mind, the language user’s or the linguist’s demonstrated that the techniques used to study polysemy in Cognitive Semantics were, in Popper’s sense, \textit{ad hoc}. Just as the approaches in Cognitive Linguistics had inherited the analytical construct of Structuralism - senses come in discrete categories - they had also inherited that the analytical technique of the Mentalist grammars – introspection can identify grammaticality. However, for Cognitive Linguistics, with no ‘ideal speaker’, the intuition of one native speaker (the linguist) cannot be assumed to represent the language. Introspection is still widely used in Cognitive Linguistics, but following the publication of Sandra & Rice’s study, descriptive semantic analysis based entirely on introspection, lost much scientific credibility within the cognitive community.

This critique of introspection as the sole basis of linguistic investigation, in combination with the establishment of Cognitive Linguistics as an independent model of language, led to the growth in experimental and corpus-
driven methods. To understand, however, the extent to which Cognitive Linguistics has always been an empirical approach to language, let us briefly revisit the foundational theories of Fillmore (1985), Talmy (1985), Langacker (1987), and Lakoff (1987).

2. Usage-based Cognitive Linguistics - unifying Mentalist and Structuralist theories, operationalising meaning and grammar

Geeraerts (2006a: 398-415, 2006b, 2007, 2010) argues that Cognitive Linguistics has recontextualised language description, returning elements that were removed by Structuralism and Mentalism-Modularism. More specifically, he shows that the tenets of the paradigm entail that meaning drives grammar. This meaning, he stresses, is conceptual, emergent, and social. Geeraerts’ idea of the socio-conceptual recontextualisation of Cognitive Linguistics is difficult to dispute, seeing the current trends. The basic argument is that there has been a linear progression from Structuralist and then Mentalist-Modularist decontextualisation to the recontextualisation of grammar in Cognitive Linguistics. However, one could perhaps go further and argue that this is but a symptom of a more fundamental unification of language theories. There are two issues here - how the theory of Cognitive Linguistics unites Mentalist and Structuralist language models, and how this unification operationalises grammar and semantics. As such, it can be said that a combination of the usage-based model and corpus-driven research resolves some of the great debates of 20th century linguistics.

2.1. Entrenchment – operationalised grammar

Langacker’s (1987) theory of entrenchment unites the two dominant 20th century models of language. If we read Langacker’s position in Chomskyan terms, the idea of an individual’s ‘competence’ is maintained. Each speaker possesses a mental grammar with intuitions about correctness in language. However, for Langacker, this individual mental grammar does not represent a language. In other words, there is no ideal speaker - each individual’s grammar is a learnt code that is unique to the extent that the individual’s exposure to language is different from other individuals’ in the speech community. It is this point that is crucial since it brings us to de Saussure’s langue. The speaker, in his or her use of language, constantly
judges a perceived set of usage norms, which in Structuralist times, would have been called *langue*.

Entrenchment, however, is more than a theory of grammaticality, it is an operationalisation of grammaticality. Langacker is primarily concerned with the status of a linguistic unit, but in the usage-based model, entrenchment can be extended to replace the notion of grammatical correctness, where the principle of frequency of use for the individual is extended to that of frequency of occurrence in the community. This operationalisation defines the phenomenon of grammar by showing how one may observe and measure it.

Every use of a structure has a positive impact on its degree of entrenchment, where [...] disuse has a negative impact. [...] Moreover, units are variably entrenched depending on the frequency of their occurrence (*driven*, for example, is more entrenched than *thrive*)... The absence of a sharp division between units and nonunits has the consequence that the scope of a grammar is not precisely delimited. (Langacker 1987: 59-60).

Here, Langacker offers an observable and quantifiable definition of an individual’s grammar and demonstrates how it relates to the grammar of a language.

The measurable definition of grammar proposed by Langacker unifies the internal Mentalist and the external Structuralist conceptions of language. This unification eliminates the need for both the *langue-parole* and competence-performance distinctions. It restores the holistic complexity of language analysis by treating the emergent structure of grammar as a result of its use by individuals in a community. Seen in this light, not only is Cognitive Linguistics recontextualising language as a social phenomenon, it is uniting the two principal 20th century models of language and defining its framework in such a way as to permit empirical research.

2.2. Categorisation - operationalised encyclopaedic semantics

Entrenchment, as a theory, bridges Mentalism and Structuralism. It is, however, a symbolic theory of grammar, based on the pairing of form and meaning. Thus, semantics is crucial and basic to any analysis that employs it. Can we also operationalise meaning? Stefanowitsch (this vol.) asks this very question and offers four operational definitions. However, before we ask that question, we must emphasise that meaning is not restricted to propositional or referential semantics in Cognitive Linguistics. Lakoff’s
Corpus-driven Cognitive Semantics

(1987) theoretical work represents the foundations of Cognitive Semantics and his position mirrors Langacker’s unification of *langue* and ‘competence’ in that it also dismantles hypothetical sub-parts of language structure, such as linguistic semantics and pragmatics. Following Fillmore’s (1985) work, Lakoff argues for a non-propositional semantics based on world knowledge – encyclopaedic semantics.

Yet, encyclopaedic semantics should not be seen as merely a combination of linguistic semantics and pragmatics; it places the holistic complexity of human experience and semiosis in the driving seat of language production. The study of language must, therefore, account for all that we know about the world from denotational reference, through cultural knowledge to even social expectations. In concrete terms, any study must simultaneously account for the semantic motivation behind and interaction between syntax, morphology, lexis, prosody, and all of this relative to discourse structures, world knowledge, and social variation. Moreover, this sea of infinite complexity varies subtly from individual to individual. It is the sum of all this that represents our object of study – meaning. The seeming impossibility of scientifically describing such multidimensional complexity is what led Structuralism to treat language use separately, the *parole*, and what led Mentalism-Modularism to dismiss it entirely. However, for a non-modular semantically driven theory of language, the complexity of language use is the basis of grammar. Can such a broad-reaching understanding of meaning be operationalised scientifically? Lakoff attempts this through the notion of conceptual categorisation.

Categorisation is a symbolic distinction between difference and similarity. This notion permits Lakoff (1987) to operationalise meaning. Things that are conceived as similar are grouped together, distinct from things that are not similar. The sets of things conceived of as similar are called concepts. This use of the notion concept is a powerful operationalisation of meaning and allows us, as linguists, to use the same analytical tool to account for the full complexity of encyclopaedic semantics. It explains why two physically different chairs are both labelled *chair*, why the lexeme *bill* is not ambiguous in a restaurant scenario, why she sneezed the napkin off the table is comprehensible, but also how conceptual metonymy and metaphor function.

The idea that, as cognitive beings, we are constantly judging similarity and dissimilarity, that through this ability we order the chaotic and dynamic environment in which we live, and that this ability and process is basic to language seems reasonable. Moreover, at a theoretical level, it is a reason-
able operationalisation of meaning. But does it work? Can this abstract definition lead to quantification and measurement? In fact, it is already the basis of much corpus-driven research. Co-occurrence and correrlation are fundamental to corpus investigation. Whether it is formal co-occurrence, such as that indentified by collocation in Collostructional Analysis, or semantic co-occurrence, such as that indentified in the usage-feature analysis of the behavioural profile approach, corpus research functions by identifying sets of similar things. This is exactly Lakoff’s understanding of conceptual categorisation. So we see that just as frequency can operationalise grammaticality, co-occurrence can operationalise categorisation.

Thus, we can say that frequency of co-occurrence, which is fundamental to corpus research, is a quantitative operationalisation of the basic theories of Cognitive Linguistics – entrenchment and categorisation. These theories, entrenchment and categorisation, explain grammar and meaning. However, co-occurrence must be understood as much more than formal co-occurrences. As stressed above, we must simultaneously account for the interaction of all dimensions of meaning in order to explain language. But it is precisely this daunting task that quantitative corpus-linguistics is best placed to achieve. Let us consider how.

2.3. Linguistic complexity - A multifactorial approach to language

We have so far established that Cognitive Linguistics has freed linguistic research of complex theoretical models, the models that were designed to tie down and render the complexity of natural language simple enough to study rigorously. We have also seen how grammar and meaning are accounted for theoretically and how these accounts are operationalised in terms that are applicable to corpus-driven research. Yet still, how is it possible to scientifically account for such an immense and complex linguistic system, a system that varies from speaker to speaker and from context to context? With no core-grammar, no langue, not even propositional or referential semantics, what hope do we have of capturing, accurately and rigorously, the conceptual structure that we believe motivates language? Is it truly possible to make generalisations and write grammars without reducing this complexity? One answer lies in multifactorial modelling. Of all the advantages of corpus-driven research, this stands out as one of the most important. Geeraerts (2006a) summarises succinctly the basic advantage:
Corpus materials provide the firm empirical basis for research, in which linguistic phenomena are statistically analysed with the help of multivariate techniques, and in which social and cultural variation is explicitly included into the multifactorial model. Geeraerts (2006a: vi)

Multivariate statistics is a powerful tool for any social scientist. Quantitative methods are essential in determining the probability that our generalisations are representative of the population but also because they help find patterns of co-occurrence that might be otherwise impossible to identify. This is especially the case when faced with so complex and multidimensional a phenomenon as language. Without language modules such as semantics, pragmatics, lexis, or syntax, language description must account for the simultaneous interaction of the different dimensions of language. Indeed, not only is multifactorial modelling of language, and the various factors that go into speech, necessary to capture the said complexity, it also represents a cognitively plausible model of language production.

Multifactorial analysis began by operationalising the various factors that impact upon the use of lexemes, morphemes, and syntactic patterns. This was done through feature analysis of large numbers of found examples. Despite the success of these first attempts (Dirven et al. 1982; Schmid 1993, Geeraerts et al. 1994; Rudzka-Ostyn 1995), manually handling the results of multifactorial feature analysis limits their interpretative power. Inspired by the research of psycholinguistics and sociolinguistics, Geeraerts et al. (1999), Gries (1999, 2001, 2003), and Speelman et al. (2003) began applying multivariate statistics to language analysis.

With the development of quantitative tools to treat corpus data and elicited data, the future looks bright. It is not difficult to imagine a future where linguists learn different analytical methods rather than study hypothetical constructs. In such a scenario, with the old dialectic between Cognitive Linguistics and its Mentalist and Structuralist precursors forgotten, linguists would no longer need to identify themselves as one theoretical camp or another. Rather, on the basis of empirical and verifiable data, linguists would dispute which method of data collection and analysis is best suited to answer a given research question. Similarly, they would question the representativity of data instead of simply choosing a different example that better suits their hypothesis. Linguists would compare results gleaned from various methods, and only then, equipped with tested hypotheses, would they advance theoretical debate. As Geeraerts (this vol.) stresses, differences in results that either confirm or falsify hypotheses, should be the only basis for theoretical discussion.
3. Quantitative corpus-driven Cognitive Semantics

Employing the usage-based model means that real language use, in all its complexity, must be the basis of linguistic research. But what does usage-based methodology entail? From corpus-extracted illustrations of a theoretical discussion to collocation studies and then confirmatory multivariate corpus-driven analysis, there is a full gamut of methodological variation (cf. Tummers, Heylen & Geeraerts 2005 and Heylen, Tummers & Geeraerts 2008 for a discussion on this point; Gilquin & Gries 2009 extend this discussion to the use of experimental data). Before quantitative corpus-driven Cognitive Semantics can become a mainstream methodological choice for linguists, it must answer two very important questions. How can we use numbers to analyse semantics and why should we do so? The problem is not corpus data or even corpus-driven analysis. Deignan (2005, 2009), for example, offers excellent examples of non-quantitative corpus-driven Cognitive Semantic research. The problematic issue is quantification. Meaning is an inherently subjective, mentally internal, and non-observable phenomenon. Why should we attempt to analyse it with quantitative techniques?

Before we can consider the strengths and weaknesses of quantitative Cognitive Semantics, we must set straight two common misconceptions about corpus methods that persist within the wider Cognitive Linguistic community.

3.1. Fallacies about corpus methodology in Cognitive Linguistics

Many linguists believe that corpus-driven research is restricted to the study of collocations and that quantitative techniques used to treat corpus data are restricted to obtaining probability values. These two beliefs are erroneous. Firstly, corpus linguistics is not restricted to the study of collocations and formal correlations. Although the study of ‘words and the company they keep’ lies at the origins of corpus linguistics, methods for studying corpus data have developed much since those times, both in terms of the kinds of questions that are asked and the kinds of techniques used to answer them. This is not to say that collocations and syntactic patterns are no longer important within the field, as we will see in the following chapters, but they are far from the only research paths possible using corpus techniques.
Secondly, quantitative techniques are not just about $p$-values. The belief that corpus linguistics is all about probability scores and that corpus linguists do not ‘get their hands dirty’ with careful close language analysis could not be further from the truth. Moreover, this is particularly false when corpus techniques are applied to semantic research. Indeed, the usage-feature or behavioural profile approach, one of the principal trends in corpus-driven lexical semantic research (Gries 2006a; Divjak 2006; Glynn 2009; Janda & Solovyev 2009; Speelman & Geeraerts 2010) and grammatical semantic research (Heylen 2005; Tummers et al. 2005; Wulff et al. 2007; Grondelaers et al. 2008; Szmrecsanyi 2010) involves the detailed and extremely laborious task of analysing a wide range of formal, semantic, and sociolinguistic features of thousands of natural language examples. Given this kind of research, arguing that corpus linguists are just interested in numbers is risible.

The quantitative treatment of data made available through meticulous manual or semi-automatic analysis, should not be seen as the hunt for $p$-values. Statistics has several roles – estimating statistical significance; detecting patterns in data; identifying the relative importance of usage factors in grammatical composition or word choice; and determining whether a proposed model or explanation accurately describes the data at hand. Probability values are but one of these roles, they indicate statistical significance. They are important because they tell us that our results would most likely be replicable, but they say no more than that. Save as confirmation that a given analysis and its interpretation are more than chance, $p$-values offer no linguistic insights *per se*.

If we accept that corpus linguistics can be more than the study of formal phenomena and that statistics is more than the search for $p$-values, then we can move on to consider the true strengths and weaknesses of quantitative corpus-driven Cognitive Semantics. These questions are left up to the authors of this volume, but let us briefly consider the main issues that lie behind many of the debates in the field.

3.2. Issues for corpus-driven semantics

3.2.1. Corpus representativity

There is, in reality, no such thing as a balanced corpus and no corpus can ever hope to be representative of a language. Corpus linguists have argued
that despite this shortcoming, a corpus is surely more representative than a single speaker. This response is not as unproblematic as it sounds, raising fundamental questions about our object of study. Modern linguistics was founded by de Saussure who delimited the object of study, *langue*, as a structured system in a given place and time. His ideas set the stage for the erroneous assumption that language is a discrete object of study, whose delimitation is not an issue. In some ways, much of the spilt ink of 20th century linguistic theory directly follows from disagreement over the object of study. From syntax and core grammar to pragmatics, linguistics argued over what language is for most of the preceding century. As stressed above, Cognitive Linguistics unifies modular language analysis with a holistic recontextualised approach to language. In light of this, one might argue that a corpus, which cannot represent the entire complexity of a language, can never be a basis for studies of language, writ large. The response is straightforward - we do not attempt to account for all of language in every study.

The usage-based model places variation, between groups and even between individuals, as an integral part of language. This we can accept *a priori*. The implications of such an assumption are brought forward when we build our corpora or extract our data for analysis. Indeed, in each corpus-driven study, the linguist is forced to answer the question: what part of language are we studying? Often practical issues dictate such choices, but in any case, they are always overtly recognised choices. These choices directly determine the scope of a study and this information should reflect variation inherent in language. Nevertheless, fine-tuning our research methods and building better, more diverse, corpora remain important keys to improving representativity. Gries (2006b; 2008) examines some of the issues at hand and the general push towards more analyses of spoken corpora is well documented in corpus linguistic journals (cf. Newman 2008a and this vol.).

It is impossible to study all of language at once (the reason why the Structuralist and Modularist programmes chose to focus on ‘parts’ of language). Corpus linguistics makes this impossibility overt and, as such, this apparent limitation is actually a blessing in disguise. It will help linguists be much more realistic about what can be scientifically said about language in a single study.
3.2.2. Negative evidence

One of the most commonly cited criticisms of corpus linguistics, a criticism that goes back to Chomsky, is that there is no negative evidence. No corpus, irrespective of its size, can possibly represent a language, let alone tell us whether a given expression or use of an expression is impossible. However, negative evidence in this sense of the term is only crucial in a rule-driven theory of language, such as that propounded by Chomsky. Since, in a usage-based model, grammaticality is based on entrenchedness, and entrenchedness is a result of use varying from individual to individual (indeed, even within the individual), there are no hard and fast rules. In a usage-based model of grammar, grammatical rules are merely generalisations about usage. For such a model of language, negative evidence is of much less importance.

To take a simple example, in a Mentalist rule-driven grammar, for the proposal of a rule that the second person copula in English takes the form are, any occurrence of *you is or *you am would disprove the rule. However, since grammaticality is a matter of degree in Cognitive Linguistics, the fact that in certain regions of England certain social groups say you is (especially in the reduced form) is incorporated into our grammar, without negating the constructional importance of the you are pattern. Instead, the use of this construction is relative to established sociolinguistic factors and is part of the system. Hence, grammatical tests and their need for negative evidence are of no concern to the usage-based linguist. Langacker’s operationalisation of grammaticality means that frequent is grammatical and infrequent is less grammatical.

For Structuralism, defining the meaning of a lexeme, such as bachelor, might be an example of the need for negative evidence. If we define a bachelor as an unmarried man, but there exists a propositional use that contradicts this meaning, such as ??a Catholic priest is a bachelor, the definition would be shown to be inadequate. In this case, one would need to redefine the lexeme (here as a “man thought of as someone who could marry” Wierzbicka 1990). The process is methodologically identical to the rule (re-)writing of Generative Grammar. If a proposed rule generates sentences that are unacceptable, then the rule is too powerful and needs to be redefined. In this way, the desire to identify the necessary and sufficient conditions for distinguishing the meaning of one lexeme from the meaning of another motivates the need for negative evidence. However, the idea of a proof, which may be falsified with a single counter example, is arguably
not informative for language description. Rather, social sciences are better served by generalisations based on samples. Given the model of language, the need for negative evidence is not so obvious for the cognitive linguist. For further discussion on this issue, within the framework of corpus Cognitive Linguistics, see Stefanowitsch (2006, 2008).

3.2.3. Frequency and salience

What can frequency tell us about language? This is a very real issue for the corpus-driven study of meaning. Although some might argue that frequency is an indicator of productivity or entrenchment, it would be difficult to claim that frequency directly equates salience. If salience is an indicator of relative semantic importance, we are faced with a problem. In terms of semantic content, most frequent often equates least semantically important, where rarity, or marked usage, indicates greater semantic importance. Givón (1991) is explicit on this point - “The marked category [...] tends to be less frequent, thus cognitively more salient, than the corresponding unmarked one.” However, in another sense of semantic importance, the reverse is sometimes true. For example, when a language more frequently refers to a given phenomenon, cultural or otherwise, we assume that it is culturally significant. This line of reasoning is basic to much of Wierzbicka (1985) and Lakoff’s (1987) research on culturally determined categories.

Salience is often cited as a crucial notion for establishing relations between categories, especially prototype structures (Durkin & Manning 1989, Geeraerts 2000). It follows that since the results of psycholinguistics are often based on relative categorisation determined by salience, understanding the relationship between frequency and salience is essential for the comparison of corpus and experimental results. Thus, although there is a relationship between frequency, salience, and semantic import, this relationship is not straightforward. Since corpus research is dependent on the study of relative frequency, how can this method be used to talk about semantic structure and how can its results be compared to experimental results?

This question may be answered by operationalising the concepts involved. It appears that there are different kinds of salience, yet how frequency is related to perceptual salience, conceptual cultural salience, or formal linguistic salience is still an open question. This debate is growing within the field and is treated by Newman (this vol.) and Schmid (this vol.),
but see also Gilquin (2006, 2010), Wiechmann (2008a, 2008b) Gilquin & Gries (2009), and Arppe et al. (2010). The answers to such questions will probably be found when concepts such as salience are more clearly defined through operationalisation. Perhaps, it will be shown that frequency can be a measure of certain kinds of salience and not others. Currently, however, the debate remains open.

3.3. Advantages of corpus-driven Cognitive Semantics

This volume should not only be seen as an attempt at identifying the limitations of a corpus-driven method, the advantages must also be developed. Let us briefly consider the basic advantages of this approach.

3.3.1. Empirical

It may seem obvious that corpus-driven research is empirical, but it is important to state this clearly because, for two different reasons, many within the research community believe that the use of corpus data does not constitute empirical research in linguistics. The first argument comes from within the experimental community and is discussed by Gries & Divjak (this vol.). Summarised briefly, the argument is that as corpus linguists, we do not have direct access to the mind, and therefore language production. It follows that we are not actually studying language production, but traces of language use extant in corpora. The response to this critique is simple - it is true. No corpus linguist studies language production. However, these traces of use are a wonderfully rich source of information on how language is used. Moreover, although we have no direct access to our object of study, nor does experimentation. Psycholinguists elicit responses to stimuli; they do not look inside the ‘black box’ itself. Thus, we collect data, they elicit it. Neither approach has direct access to the mind or its functioning. Psycholinguistic experimentation is better placed to look at processing and dynamic issues in language production, while corpus linguistics is better placed to look at natural use and, perhaps, even the intersubjective dimension of language (cf. Glynn & Krawczak submitted). Although, as stressed by Newman (2008a; this vol.), there is a terrible lack of spoken and interactive corpora for exploring these dimensions of language.
The second argument comes from the community that uses introspection to perform linguistic analysis. Their argument is reasonable - since our theory of language is a semantics-driven theory of language and semantics is necessarily subjective, that is internal to the mind, introspection is the only viable method for its analysis. Talmy (2008: xix) is explicit about this point. He stresses that corpus research “cannot directly yield many abstract linguistic patterns”. His argument is similar to that of the psycholinguists’, save, at least, that he is correct in saying that introspection does access the mind (unlike in psycholinguistic experimentation). His point is not disputed but we have two important additions to the discussion. Firstly, despite this limitation, the patterns of natural language usage are an incredibly rich source for working out how people use language and this can produce a very good picture of language structure. Secondly, as stressed by Geeraerts (this vol.), introspection is vital to corpus linguistics. The categories chosen for study, the actual analysis of those categories, and, of course, the hypotheses that the study will test are all a result of introspection. It is not that empirical research replaces introspection, rather introspection is used to propose hypotheses, which then need to be tested by operationalising the questions and designing a study that will adequately answer those questions. If we cannot find a means to adequately answer a scientific question, then we are in the same position as scientists in all fields and must continue to look for a means to do so. This is no argument to remain at the hypothetical stage of enquiry, based entirely upon introspection.

3.3.2. Quantitative

If one has empirical data, then one may quantify them. However, it is far from self-evident that corpus-driven semantic research should be quantitative. Since debating this proposal is the point of the current volume, we will not delve too deeply into this discussion. However, the basic advantage offered by a quantitative approach to semantics is that it permits the operationalisation of our studies. The benefits of, and indeed need for, operationalisation are discussed by Stefanowitsch (this vol.). The questions, of course, are – can and should we operationalise semantics quantitatively? These questions are specifically broached by Newman (this vol.) and Schmid (this vol.) but, to a greater or lesser extent, they are considered by all the contributions to this volume. One basic point should be established - no one wishes to argue that we should reduce semantic questions to num-
bers. It is not the frequency _per se_ of linguistic features that is of interest, but what this says about usage, the relative association of forms and meanings in context.

Even if quantification is possible, the next question is - why do we need it, what do we gain from employing quantitative techniques? Firstly, they permit, and indeed encourage, the empirical cycle, explained by Geeraerts (this vol.). It is much easier for the research community to verify the results of a study when one can see exactly how those results were obtained. Secondly, quantified data can be examined using statistics. Statistical analysis of data offers several advantages: (i) Confirmatory statistics allows one to determine the statistical significance of the results of an analysis. In other words, what is the probability that similar results would be obtained if further examples were examined in the same way? (ii) Multivariate statistics allows one to identify patterns in usage that would be effectively impossible to identify using introspection. (iii) Statistical modelling allows one to test the accuracy of an analysis. This kind of measure determines how much of the variation, in a given sample, a given analysis can explain. These multivariate techniques do not only examine the effects of numerous factors of usage simultaneously, they include the interaction of these factors as well. Thus, Empiricism permits quantification and quantification permits statistical confirmation but also multifactorial analysis. This is the third basic advantage.

3.3.3. _Multifactorial_

Empirical research permits quantification, and, in turn, quantification permits multifactorial analysis. Although this may seem like a narrow addition to the discussion, it is arguably the most important facet of corpus-driven semantic research. As stressed in section 2.3, it is the complexity of our object of study that is our greatest hurdle. The cognitive model of language insists that we simultaneously take into account the full socio-cognitive spectrum of factors that influence and motivate language. Whether the research is based on questionnaires, advanced eye-tracking experiments, word space models derived from mass computation of huge corpora, or the nitty-gritty manual analysis of semantic features of a small corpus sample, multivariate modelling is a basic and crucial tool. It is almost impossible to imagine how one might account for all the factors, and the interaction of these factors, that go into language production and comprehension. How-
ever, should it ever be possible, it will only be so with the aid of such analyti-
cal tools. Therefore, arguably, it is precisely the possibility of multifac-
torial analysis that makes quantitative empirical research, be that corpus-
driven or experimental, essential for linguistic semantics.

4. Cognitive Semantics – lexical and grammatical meaning

This decade has seen a veritable surge of corpus-driven Cognitive Lin-
guistic research. This volume can only scratch the surface of the theoretical
questions that such a method raises and offer a small sample of the research
it makes possible. The focus of the volume is specifically the application of
quantitative corpus-driven methods to Cognitive Semantics. Yet, even lim-
iting the horizons in this manner, the possible applications and issues go
beyond the purview of any single volume. Let us place the work presented
here in its methodological context. This will allow the reader to better
understand its importance, but also see where future directions of the field
lie.

4.1. Survey of the field

4.1.1. Semasiology–onomasiology and semantic schematicity

For Cognitive Semantics, one has two possible objects of study. Firstly, one
may ask what forms are available to express a given concept (onomasiol-
ogy or synonymy). Secondly, one can ask what concepts are expressed by a
given form (semasiology or polysemy). Having categorised the research as
onomasiological or semasiological, one may further divide it into the study
of schematic or non-schematic form and meaning. Schematic meanings are
more typically expressed by grammatical forms, such as morphemes and
grammatical constructions while non-schematic meanings are more typi-
cally expressed by lexical forms. The semasiological–onomasiological
division crosses both objects of study. For grammatical meaning, the ono-
masiological study of schematic concepts includes, for example, syntactic
alternations or the choice between grammatical cases. The semasiological
study of grammatical forms looks more closely at a given construction or
morpheme. The same distinction applies to lexical semantics, effectively
drawing a distinction between the study of polysemy networks and near-
synonymy lexical fields. This leads to four research questions – (i) What words do people choose? (ii) What grammatical constructions do people choose? (iii) How do people use a word? (iv) How do people use a grammatical construction? We can label these research questions respectively as lexical synonymy, grammatical synonymy, lexical polysemy, and grammatical polysemy.

Although there is no clear divide between lexical and grammatical forms and obviously even less so between the schematic and non-schematic meaning, the onomasiological–semasiological division is reasonably clear-cut. This ‘form-first’ or ‘concept-first’ distinction is arguably fundamental to semantic research (Geeraerts et al. 1994; Geeraerts 2006a). However, even here, Glynn (2010) argues that there exists a continuum between these two dimensions, depending on the granularity of the study. This is because all variation in prosody, syntax, and morphology is technically a change in form that can represent a subtle change in meaning. To this extent, the line between semasiology and onomasiology is also not ultimately discernable. Newman (2008b; this vol.) broaches this question in his discussion on the problems of studying the lemma, arguably a very coarse-grained level of analysis. Nevertheless, a methodological division does exist, which demands that research begin with either a concept – examining the various forms that express it, or with a form – examining the various meanings expressed by it.

The four research domains listed above can be used to categorise all research in Cognitive Semantics. The work on culturally determined concepts, semantic frames, cognitive models, and metaphors (Wierzbicka 1985, Kövecses 1986, Lakoff 1987: case study 1) is obviously lexical onomasiology, just as the research in polysemy is typically lexical semasiology (Lakoff 1987: case study 2, Fillmore & Atkins 1992, Cuyckens 1995). The studies in syntactic alternations, popular in Construction Grammar, are clearly grammatical onomasiology. Similarly, Langacker (2000) and Talmy (2000) examine the differences between various grammatical profilings, an inherently schematic-semantic, morpho-syntactic, and onomasiological domain of research. This leaves the semasiological study of individual grammatical forms and syntactic patterns. Such research is less well-known but equally represented. Important studies include Lakoff (1987: case study 3), Janda (1993), Goldberg (1995), Rudzka-Ostyn (1996), Dąbrowska (1997), Geeraerts (1998).

Having established the four research domains of Cognitive Semantics, we need to introduce two methodological variables specific to corpus-
driven analysis - the relative use of observable objective linguistic features and the relative use of statistical techniques.

4.1.2. Hard statistics–soft statistics and objectivity

Methodologically, corpus-driven research in semantics is far from being a single established approach. There are different ways of collecting data, analysing the data, and then a wide range of quantitative techniques available for treating the results of these analyses.

First, research in a quantitative semantic study can be restricted to the analysis of formal observable phenomena or it can include semantic non-observable phenomena. This is important since the inclusion of non-observable, and therefore subjectively determined, factors in semantic analysis greatly reduces the objectivity of the study. Moreover, if the analysis is restricted to formal observable characteristics, one has the possibility of the automatic treatment of data. It follows that if the analysis is automatic, one may examine large quantities of examples. As corpora increase in size, and as methods for automatically treating them improve, new possibilities for this line of research are beginning to emerge. Examples of automatic analyses that seek to answer semantic questions include the collostructional analyses of Stefanowitsch & Gries (2003), Gries & Stefanowitsch (2004) and Hilpert (2008) and the lexical onomasiological studies of Arppe (2008) and Levshina et al. (forthc.). However, the application of computational techniques, such as word space modelling, to Cognitive Semantic research is also emerging. The Sem•metrix project, led by D. Geeraerts, is an example of such an approach (Peirsman et al. 2010).

The lack of semantic richness in formal automatic analysis is offset by the fact that its results can be more reliable, not only due to the objective nature of the analysis, but because a larger sample improves representativity. Nevertheless, despite their merit, such approaches arguably miss much of what is important for semantic research. It is questionable whether linguists will ever be able to sufficiently describe language structure based exclusively upon formal patterns of usage. In light of this, small-scale detailed manual semantic analysis, typical of the usage-feature behavioural profile approach, remains an important line of investigation. Both lines of research, as well as their pros and cons, are treated in this volume.

Determining the degree of statistical sophistication for the treatment of data is not a straightforward question. With small quantities of data, due to
limited resources imposed by transcribing spoken language, scarcity of historical sources, or the practical constraints of fieldwork, one can do little more than count and compare raw numbers. Moreover, the meticulous manual analysis of usage features considerably restricts the number of examples in a dataset, in turn, restricting the efficacy of more advanced statistical techniques. Finally, the technical competence of the author is a real, if not scientific, factor that influences the choice of statistical methods. At least for the current generations of linguists, statistical training is rare, though this is set to change in the near future.

Quantitative approaches to semantic structure within Cognitive Linguistics begin with counting examples of a certain kind and comparing them to examples of another kind (e.g.: Hanegreefs 2004; Davidse et al. 2008; Dziwirek & Lewandowska-Tomaszczyk 2009). Such quantitative data may employ tests for statistical significance, such as the t-Test or Chi-square test. Other quantitative research employs exploratory techniques such as Cluster Analysis and Correspondence Analysis (Gries 2006a, Divjak 2006, Szeld & Geeraerts 2008, Glynn 2009). These techniques may also be combined with significance testing, but their focus is identifying patterns and associations in the data. Finally, confirmatory statistics such as Linear Discriminant Analysis and Logistic Regression are complex, yet extremely powerful techniques (Gries 2003, Heylen 2005, Speelman and Geeraerts 2010). Confirmatory statistical modelling presents many possibilities for the multivariate description of language, as well as the verification of that description. Glynn & Robinson (in press) offers a survey of the statistical approaches popular in Cognitive Semantics.

4.1.3. Social dimensions of semantics

Special mention must be made of the sociolinguistic element in corpus-driven research. This line of research is of growing importance (Geeraerts 2005, Croft 2009, Geeraerts & Speelman 2010) and is especially significant to quantitative corpus-driven approaches for two reasons. Firstly, sociolinguistics, along with psycholinguistics, has employed some of the most advanced statistics for language study and has, to some extent, paved the way forward for Cognitive Semantics. Secondly, it is precisely this element of language structure that is effectively impossible to account for using introspection and difficult to adequately account for using experimentation. It is for these reasons that the interaction of meaning and society is at the heart
of corpus-driven Cognitive Semantic research. Two recent anthologies, Kristiansen & Dirven (2008), Geeraerts et al. (2010), are devoted to the subject.

4.2. Case studies in the field

We can now consider the field of research. Tables 1 and 2 were compiled by collecting the relevant studies from a range of sources. These sources include the anthologies listed in footnote 2, four Cognitive Linguistics journals (Cognitive Linguistics, Annual Review of Cognitive Linguistics, Language and Cognition and Constructions and Frames), and three important corpus linguistics journals (Corpus Linguistics and Linguistic Theory, Corpora, International Journal of Corpus Linguistics). Tables 1 and 2 list both the object of study and the method employed. The methodological information offered includes the degree of statistical complexity: counts, collocation (including Collostructional Analysis), exploratory statistics (Hierarchical Cluster Analysis HCA; Principal Component Analysis PCA; Configural Frequency Analysis CFA; Multiple Correspondence Analysis MCA; Multidimensional Scaling MDS), and confirmatory multifactorial analysis (Linear Discriminant Analysis LDA; Profile-Based Analysis PBA; Logistic Regression Analysis LRA; Mixed Effects Logistic Regression MER). These techniques are explained by various authors in Glynn and Robinson (in press). When more than one technique is employed, only the most advanced is listed. The degree of objectivity is also indicated – studies based entirely on formal analysis of observable data versus studies that include at least some non-observable semantic usage-features. The tables also indicate if a study includes extralinguistic dimensions – socio-linguistic, diachronic, or contrastive factors. The studies are sorted chronologically, relative to whether they have a semasiological or onomasiological emphasis and whether they treat lexical or grammatical forms and concepts.
Table 1. Corpus-Driven Cognitive Semantics - Lexical Synonymy and Polysemy

<table>
<thead>
<tr>
<th>Object</th>
<th>Method</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lexical Onomasiological</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbs of informing (Eng.)</td>
<td>-</td>
<td>counts</td>
</tr>
<tr>
<td>Nouns for clothing (Dch.)</td>
<td>socio.</td>
<td>counts</td>
</tr>
<tr>
<td>Verbs of starting (Eng.)</td>
<td>-</td>
<td>counts</td>
</tr>
<tr>
<td>Verbs of response (Eng.)</td>
<td>diach.</td>
<td>counts</td>
</tr>
<tr>
<td>Terms for football (Dch.)</td>
<td>socio.</td>
<td>conf. PBA</td>
</tr>
<tr>
<td>Abstract nouns (Eng.)</td>
<td>-</td>
<td>counts</td>
</tr>
<tr>
<td>Verbs of posture (Eng.)</td>
<td>-</td>
<td>counts</td>
</tr>
<tr>
<td>Concept of ANGER (Eng.)</td>
<td>diach.</td>
<td>counts</td>
</tr>
<tr>
<td>Verbs of eating (Eng.)</td>
<td>-</td>
<td>counts</td>
</tr>
<tr>
<td>Verbs of intention (Rus.)</td>
<td>-</td>
<td>explor. HCA</td>
</tr>
<tr>
<td>Verbs of attempting (Rus.)</td>
<td>-</td>
<td>explor. HCA</td>
</tr>
<tr>
<td>Verbs of becoming (Spanish)</td>
<td>-</td>
<td>explor. MDS</td>
</tr>
<tr>
<td>Verbs of separation (Dch., Eng., Germ., Swed.)</td>
<td>contr.</td>
<td>explor. HCA</td>
</tr>
<tr>
<td>Verbs of cognition (Fin.)</td>
<td>-</td>
<td>conf. LRA</td>
</tr>
<tr>
<td>Concepts POSITIVE, NEGATIVE (Hungarian)</td>
<td>socio.</td>
<td>explor. MCA</td>
</tr>
<tr>
<td>Verbs of starting (Eng., Rus.)</td>
<td>contr.</td>
<td>explor. HCA</td>
</tr>
<tr>
<td>Verbs of reception (Dch.)</td>
<td>diach.</td>
<td>counts</td>
</tr>
<tr>
<td>Adverbs for ‘again’ (Mnd.)</td>
<td>-</td>
<td>conf. CFA</td>
</tr>
<tr>
<td>Concepts HAPPY, SAD (Rus.)</td>
<td>-</td>
<td>explor. HCA</td>
</tr>
<tr>
<td>Concepts LOVE, HATE (Pol.)</td>
<td>contr.</td>
<td>colloc. count</td>
</tr>
<tr>
<td>Verbs of posture (Eng.)</td>
<td>diach.</td>
<td>colloc. count</td>
</tr>
<tr>
<td>Terms for OTHER (Eng.)</td>
<td>socio.</td>
<td>conf. LRA</td>
</tr>
<tr>
<td>Terms for FOOTBALL (Port.)</td>
<td>socio.</td>
<td>conf. PBA</td>
</tr>
<tr>
<td>Verbs of causation (Dch.)</td>
<td>socio.</td>
<td>conf. LRA</td>
</tr>
<tr>
<td>Verbs <em>come, go</em> (Jap., Kor.)</td>
<td>contr.</td>
<td>counts</td>
</tr>
<tr>
<td>Verbs of bothering (Eng.)</td>
<td>socio.</td>
<td>conf. MER</td>
</tr>
<tr>
<td>Verbs of cognition (Pol.)</td>
<td>socio.</td>
<td>conf. LRA</td>
</tr>
<tr>
<td>Verbs of causation (Dch.)</td>
<td>socio.</td>
<td>conf. MER</td>
</tr>
<tr>
<td>Verbs of possibility (Eng., Fr.)</td>
<td>contr.</td>
<td>conf. LRA</td>
</tr>
<tr>
<td>Concept LIBERTY (Eng.)</td>
<td>socio.</td>
<td>conf. LRA</td>
</tr>
<tr>
<td><strong>Lexical Semasiological</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verb <em>ask</em> (Eng.)</td>
<td>diach.</td>
<td>counts</td>
</tr>
<tr>
<td>Verb <em>need to</em> (Eng.)</td>
<td>socio.</td>
<td>colloc. count</td>
</tr>
<tr>
<td>Verb <em>run</em> (Eng.)</td>
<td>-</td>
<td>explor. HCA</td>
</tr>
<tr>
<td>Verb, Noun <em>hassle</em> (Eng.)</td>
<td>socio.</td>
<td>explor. MCA</td>
</tr>
<tr>
<td>Verb, Noun <em>bother</em> (Eng.)</td>
<td>socio.</td>
<td>conf. LRA</td>
</tr>
<tr>
<td>Adjective <em>deep</em> (Germ.)</td>
<td>-</td>
<td>colloc.</td>
</tr>
<tr>
<td>Verb <em>annoy</em> (Eng.)</td>
<td>socio.</td>
<td>explor. MCA</td>
</tr>
</tbody>
</table>

Legend: Adj. - Adjective; Adv. - Adverb; Alt. - Alternation; Cx. - Construction; Imperf. - Imperfective; Perf. - Perfective; Dch. - Dutch; Eng. - English; Fin. - Finnish; Fr. - French; Germ. - German; Jap. - Japanese; Kor. - Korean; Mnd. - Mandarin; Pol. - Polish; Port. - Portuguese; Rus. - Russian; Swed. - Swedish; Ukr. - Ukrainian; contr. - contrastive; diach. - diachronic; socio. - sociolinguistic; confirm. - confirmatory statistics; explor. - exploratory statistics; sem - semantic annotation, form - formal annotation.
<table>
<thead>
<tr>
<th>Object</th>
<th>Method</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alt. Phrasal Verb syntax (Eng.)</td>
<td>- counts</td>
<td>form Gries 1999</td>
</tr>
<tr>
<td>Alt. Future Cxs (Eng.)</td>
<td>- counts</td>
<td>form Szmrecsanyi 2003</td>
</tr>
<tr>
<td>Alt. Adj. Syntax (Eng.)</td>
<td>- confirm. LDA</td>
<td>sem Wulff 2003</td>
</tr>
<tr>
<td>Alt. Adj. Inflection (Eng.)</td>
<td>- explor. PCA</td>
<td>form Gries 2003</td>
</tr>
<tr>
<td>Aspect Prepositions (Eng.)</td>
<td>- counts</td>
<td>form Newman &amp; Rice 2004b</td>
</tr>
<tr>
<td>Alt. Adj. Inflection (Dch.)</td>
<td>socio. confirm. LRA</td>
<td>form Tummers et al. 2005</td>
</tr>
<tr>
<td>Alt. Middle Field Syntax (Germ.)</td>
<td>socio. confirm. LRA</td>
<td>sem Heylen 2005</td>
</tr>
<tr>
<td>Alt. go V–go and V Cxs (Eng.)</td>
<td>- colloc. collostr.</td>
<td>form Wulff 2006</td>
</tr>
<tr>
<td>Grammatical Cases in Slavic</td>
<td>contr. explor. MDA</td>
<td>sem Clancy 2006</td>
</tr>
<tr>
<td>Alt. Dative</td>
<td>socio. confirm. LRA</td>
<td>sem Bresnan et al. 2007</td>
</tr>
<tr>
<td>Alt. Presentative Cxs (Dch.)</td>
<td>socio. confirm. LRA</td>
<td>sem Grondelaers et al. 2007, 2008</td>
</tr>
<tr>
<td>Alt. Nominal – Clausal Compl (Eng.)</td>
<td>- colloc. collostr.</td>
<td>form Wiechmann 2008b</td>
</tr>
<tr>
<td>Alt. Phrasal Verb syntax (Eng.)</td>
<td>socio. confirm. LRA</td>
<td>sem De Sutter 2009; et al. 2008</td>
</tr>
<tr>
<td>Alt. Genitive Cxs (Eng.)</td>
<td>socio. confirm. LRA</td>
<td>sem Szmrecsanyi 2010</td>
</tr>
<tr>
<td>Alt. Imperf.–Perf. Cxs (Rus.)</td>
<td>- confirm. MER</td>
<td>form Divjak this vol.</td>
</tr>
<tr>
<td>Alt. Case – Adposition (Estonian)</td>
<td>socio. confirm. LRA</td>
<td>sem Klavan in press</td>
</tr>
<tr>
<td>Grammatical Semasiological</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case Dative (Pol.)</td>
<td>- counts</td>
<td>sem Rudzka-Ostyn 1996</td>
</tr>
<tr>
<td>Cxs. think nothing of Gerund; Imper.; Progr.; Ditrans. (Eng.)</td>
<td>- collocol. collostr.</td>
<td>form Stefanowitsch &amp; Gries 2003</td>
</tr>
<tr>
<td>Cx. into-Causative (Eng.)</td>
<td>- confirm. CFA</td>
<td>form Gries &amp; Stefanowitsch 2004</td>
</tr>
<tr>
<td>Cx. into-Causative (Eng.)</td>
<td>socio. collocol. collostr.</td>
<td>form Wulff et al. 2007</td>
</tr>
<tr>
<td>Cxs Future (Eng., Dan., Swed., Germ.)</td>
<td>diach. collocol. collostr.</td>
<td>form Hilpert 2008a</td>
</tr>
<tr>
<td>Cx. Comparative (Eng.)</td>
<td>- collocol. collostr.</td>
<td>sem Hilpert 2008b</td>
</tr>
<tr>
<td>Aux. shall; Aspect Perfect (Eng.)</td>
<td>diach. explor. HCA</td>
<td>form Gries &amp; Hilpert 2008</td>
</tr>
<tr>
<td>Cx. good PrP (Eng.)</td>
<td>- counts</td>
<td>sem Zeschel 2009</td>
</tr>
<tr>
<td>Cx. mit-Predicative (Germ.)</td>
<td>- confirm. CFA</td>
<td>form Hilpert 2009</td>
</tr>
<tr>
<td>Cx. Dative (Dch.)</td>
<td>diach. collocol. collostr.</td>
<td>form Colleman 2009, this vol.</td>
</tr>
<tr>
<td>Cx. Raising (Eng., Dch.)</td>
<td>diach. collocol. collostr.</td>
<td>form Noël &amp; Colleman 2010</td>
</tr>
<tr>
<td>Cx. Verb Poss. way (Eng.)</td>
<td>- explor. HCA</td>
<td>form Gries &amp; Stefanowitsch 2010</td>
</tr>
<tr>
<td>Cx. Benefactive Ditransitive (Eng.)</td>
<td>socio. counts</td>
<td>form Colleman 2010</td>
</tr>
<tr>
<td>Cx. V – NP (Eng.)</td>
<td>- collocol. collostr.</td>
<td>form Wulff this vol.</td>
</tr>
<tr>
<td>Cx. Gerund Compl. Clause (Eng.)</td>
<td>- collocol. collostr.</td>
<td>form Hilpert this vol.</td>
</tr>
<tr>
<td>Cx. for Durative Adv. (Eng.)</td>
<td>- collocol. collostr.</td>
<td>form Fuhs this vol.</td>
</tr>
<tr>
<td>Suffix dissposessive –ont (Dch.)</td>
<td>diach. counts</td>
<td>sem Delorge &amp; Colleman in press</td>
</tr>
<tr>
<td>Cx. Epistemic stance (Eng.)</td>
<td>socio confirm. MER</td>
<td>sem Glynn &amp; Krawczak submitted</td>
</tr>
</tbody>
</table>
The first thing one will notice looking at the tables is that for each object of study, there is a tendency towards more sophisticated quantitative techniques over time. Just as noticeable is the tendency for semantic analysis in lexical research contrasted by formal analysis in grammatical research. Although this is in part due to a large amount of collocation studies of grammatical constructions, there is no reason to suppose that collocation should not be equally applied to the study of lexical semantics. As Zeschel (this vol.) shows, the collocation tradition, which was originally lexically orientated, still has a great deal to offer the study of lexical semantics. Noteworthy as well is a small increase in the amount of studies including sociolinguistic, diachronic, and contrastive parameters in their research.

The tables also reveal the biases in research interests across the research paradigm. In lexical research, verbs obviously have an over representation, nouns, adjectives, and adverbs being poorly represented. Also, and somewhat surprisingly, seeing the history of Cognitive Semantics, prepositions are entirely absent. Some of this bias is because verbs, due to their argument structure and its participants, offer more observable features for analysis, facilitating quantitative study. Overcoming the difficulties of studying other parts of speech should, nevertheless, be a goal in the research field. Equally noticeable is the emphasis upon onomasiological research. Again, there is a likely methodological reason. It is relatively easy to distinguish the use of forms compared to distinguishing uses, especially since formal and extralinguistic factors may often contribute a lot to explaining onomasiological variation. The research field needs to approach the difficulties of semasiological lexical description and this volume offers two contributions, Zeschel (this vol.) and Glynn (this vol.), that attempt this.

The grammatical research also shows clear biases. Obviously, the popularity of the Collostructional Analysis accounts for much of this. However, the fact that effectively all onomasiological research examines syntactic alternations and all semasiological research looks at syntactic-lexical pairing is a clear and serious shortcoming. Although there are some important exceptions to this trend, such as the work on adjectival inflections and aspectual categories, the field must move towards other kinds of grammatical semantics. The semantics encoded by phonological structures, such a prosody, are entirely absent. The lack of research in grammatical case is especially surprising, seeing the rich tradition of this in Cognitive Linguistics. This volume corrects this imbalance with Divjak’s (this vol.) research on aspect. However, both the semasiological and onomasiological research in grammatical semantics needs to broaden its field of research.
Perhaps, the most striking gap in the research is the lack of work on discourse questions and broad culturally-determined concepts, such as those proposed by Lakoff (1987). The first steps have been taken in the study of Idealised Cognitive Models (Dziwirek & Lewandowska 2009, Glynn forthc.), but this volume does not offer any such research. Extending this method to more culturally-orientated concepts is a clear goal for future research. The study of discourse is difficult due to the lack of adequate spoken corpora. Fischer (this vol.) offers an important contribution on this front by developing a corpus especially designed to capture such discourse structures. As the field matures, such lines of inquiry will certainly attract more attention. The next step is to develop the methods to answer broader semantic research questions. This volume makes a small, but important step, in this direction.

Notes

1. I would like to thank G. Gilquin, M. Hilpert, and C. Paradis for their help. All shortcomings remain my own.

2. Testimony to the surge in corpus-driven research is the large number of anthologies, devoted entirely or substantially, to corpus-based Cognitive Linguistics. A list of these includes Gries & Stefanowitsch (2006), Stefanowitsch & Gries (2006), Gonzalez-Marquez et al. (2007), Kristiansen & Dirven (2008), Zeschel (2008), Lewandowska-Tomaszczyk & Dziwirek (2009), Gilquin (2009), Gries et al. (2009). Geeraerts et al. (2010), Rice & Newman (2010), Marzo et al. (2010), Glynn & Robinson (in press), and Gries & Divjak (forthc.). Geeraerts (2006b) offers a detailed description of the generalisation of this trend across the field.

3. R. Gibbs and M. Tomasello lead the field. One cannot even begin to list the most important contributions here, but early studies include Gibbs (1990) and Tomasello (2000).

4. Although one may interpret Langacker’s position in this way, from a strictly cognitive perspective, the idea of competence is not at all relevant. Cf. Paradis (2003) for a discussion on the notion of competence.

5. The use of frequency as a means for operationalising grammar deserves discussion beyond the scope of this introduction. Stefanowitsch (this vol.) shows its importance at a methodological level and Bybee (2007), amongst others, has fully integrated this notion into her research.
6. Langacker (1987: 100) is explicit about the relationship between entrenchment and frequency. Schmid (2000: 39) is still more explicit, proposing, what he terms, the Corpus-to-Cognition Principle – “Frequency in text instantiates entrenchment in the cognitive system.” Schmid has since, however, distanced himself from this position.


8. For sake of brevity, recent monographs in the field have not been included. For the reader’s reference, books devoted to the field include Geeraerts et al. (1999), Schmid (2000), Fischer (2000), Gries (2003), Mukherjee (2005), Szmrecsanyi (2006), Hilpert (2008), Arppe (2008), Wulff (2009a), Divjak (2010), and Gilquin (2010).


References

Arppe, Antti
2008 *Univariate, Bivariate and Multivariate Methods in Corpus-Based Lexicography – A study of synonymy*. Helsinki: University of Helsinki.


Arppe, Antti and Juhani Järvikivi

Arppe, Antti, Gaëtanelle Gilquin, Dylan Glynn, Martin Hilpert, Arne Zeschel
Baayen, R. Harald  


Bresnan, Joan, Anna Cueni, Tatiana Nikitina and Harald Baayen  

Bybee, Joan  

Bybee, Joan and David Eddington  

Clancy, Steven J.  

Colleman, Timothy  
2009 The semantic range of the Dutch double object construction: A collocational perspective. Constructions and Frames 1: 190-221.


Croft, William  

Cuyckens, Hubert  

Dąbrowska, Ewa  

Davidse, Kristin, Lieselotte Brems and Liesbeth De Smedt  
De Sutter, Gert  

De Sutter, Gert, Dirk Speelman and Dirk Geeraerts  

Deignan, Alice  


Delorge, Martine  

Delorge, Martine and Timothy Colleman  

Deshors, Sandra and Stefan Th. Gries  
in press  A case for the multifactorial assessment of learner language: The uses of *may* and *can* in French-English interlanguage. In *Polysemy and Synonymy. Corpus methods and applications in Cognitive Linguistics*, Dylan Glynn and Justyna Robinson (eds.). Amsterdam / Philadelphia: John Benjamins.

Dirven, René and John Taylor  

Dirven, René, Louis Goossens, Yvan Putseys and Emma Vorlat  
Diessel, Holger

Divjak, Dagmar

Divjak, Dagmar and Nick Fieller

Divjak, Dagmar and Stefan Th. Gries

Durkin, Kevin and Jocelyn Manning

Dziwirek, Katarzyna and Barbara Lewandowska-Tomaszczyk

Fabiszak, Małgorzata, Anna Hebda, Iwona Kokorniak and Karolina Krawczak
in press The prefix semantics and grammar in the Polish verb ‘to think’. In *Polysemy and Synonymy. Corpus methods and applications in Cognitive Linguistics*, Dylan Glynn and Justyna Robinson (eds.). Amsterdam / Philadelphia: John Benjamins.

Fillmore, Charles
Fillmore, Charles and Beryl Atkins

Fischer, Kerstin

Geeraerts, Dirk
Geeraerts, Dirk and Speelman, Dirk

Geeraerts, Dirk, Gitte Kristiansen and Yves Peirsman (eds.)

Geeraerts, Dirk, Stefan Grondelaers and Dirk Speelman
1999  *Convergentie en Divergentie in de Nederlandse Woordenschat*. Amsterdam: Meertens Instituut.

Geeraerts, Dirk, Stefan Grondelaers and Peter Bakema

Gevaert, Caroline

Gibbs, Raymond

Gilquin, Gaëtanelle

Gilquin, Gaëtanelle (ed.)

Gilquin, Gaëtanelle and Stefan Th. Gries

Givón, Talmy
Glynn, Dylan


Glynn, Dylan and Justyna Robinson (eds.)
in press Polysemy and Synonymy. Corpus methods and applications in Cognitive Linguistics. Amsterdam / Philadelphia: John Benjamins.

Glynn, Dylan and Karolina Krawczak
submitted Social cognition, Cognitive Grammar and corpora. A multifactorial approach to subjectification. Language and Cognition

Goldberg, Adele

Gonzalez-Marquez, Monica, Irene Mittleberg, Seanna Coulson and Michael Spivey (eds.)

Goossens, Louis
Gries, Stefan Th.


Gries, Stefan Th. and Anatol Stefanowitsch (eds.)


Gries, Stefan Th. and Anatol Stefanowitsch


Gries, Stefan Th. and Dagmar Divjak (eds.)


Gries, Stefan Th. and Martin Hilpert


Gries, Stefan Th., Beate Hampe and Doris Schöenefeld


Gries, Stefan Th., Stefanie Wulff and Mark Davies (eds.)  
2009 *Corpus Linguistic Applications: Current studies, new directions.* Amsterdam: Rodopi.

Grondelaers, Stefan, Dirk Geeraerts and Dirk Speelman  

Grondelaers, Stefan, Dirk Speelman and Dirk Geeraerts  

Hanegraeufs, Hilde  

Heylen, Kris  

Heylen, Kris, José Tummers and Dirk Geeraerts  

Hilpert, Martin  


Janda, Laura

Janda, Laura and Valery Solovyev


Jing-Schmidt, Zhuo and Stefan Th. Gries

Jones, Steven, Carita Paradis, Lynne Murphy and Caroline Willners

Kabata, Kaori and Jeong-Hwa Lee

Klavan, Jane

Kövecses, Zoltán

Kristiansen, Gitte and René Dirven (eds.)

Lakoff, George

Langacker, Ronald  

Lehrer, Adrienne  

Levshina, Natalia, Dirk Speelman and Dirk Geeraerts  

Lewandowska-Tomaszczyk, Barbara and Katarzyna Dziwirek (eds.)  

Majid, Asifa and Melissa Bowerman  

Marzo, Stefania, Kris Heylen and Gert De Sutter (eds.)  

Mukherjee, Joybrato  

Newman, John  

Newman, John and Sally Rice  


Noël, Dirk and Timothy Colleman

Nokkonen, Soili

Paradis, Carita

Peirsman, Yves, Kris Heylen and Dirk Geeraerts

Rice, Sally and John Newman (eds.)

Rudzka-Ostyn, Brygida


Sandra, Dominiek and Sally Rice  
1995  Network analysis of prepositional meaning: mirroring whose mind - the linguist’s or the language user’s? *Cognitive Linguistics* 6: 89-130.

Schmid, Hans-Jörg  


Schulze, Rainer  

Soares da Silva, Augusto  

Speelman, Dirk  

Speelman, Dirk and Dirk Geeraerts  

Speelman, Dirk, José Tummers and Dirk Geeraerts  

Speelman, Dirk, Stefan Grondelaers and Dirk Geeraerts  

Stefanowitsch, Anatol  

Stefanowitsch, Anatol and Stefan Th. Gries

Stefanowitsch, Anatol and Stefan Th. Gries (eds.)
2006 *Corpus-Based Approaches to Metaphor and Metonymy*. Berlin / New York: Mouton de Gruyter.

Szélid, Veronika and Dirk Geeraerts

Szmrecsanyi, Benedikt

Talmy, Leonard

Tomasello, Michael

Tummers, José, Dirk Speelman and Dirk Geeraerts

Tummers, Jose, Kris Heylen and Dirk Geeraerts
Wiechmann, Daniel

Wierzbicka, Anna

Wulff, Stefanie

Wulff, Stefanie, Anatol Stefanowitsch and Stefan Th. Gries

Zeschel, Arne (ed.)

Zeschel, Arne