Polysemy and synonymy
Cognitive theory and corpus method

Dylan Glynn
Lund University

1. Introduction. Theory and Method

The idea of ‘corpus semantics’, just like the possibility of ‘quantifying meaning’, is not self-evident. This introduction to the field of corpus-driven Cognitive Semantics attempts to explain how semantic analysis can, and indeed, should turn to corpus methods. It also explains why quantitative techniques are needed in this endeavour.

How can we identify and explain the semantic structuring of language empirically? The usage-based approaches of post-Generativist and post-Structuralist linguistics avoid positing analytical constructs that form the basis of language structure. However, without a structurally independent langue or an ‘ideal’ speaker’s competence, against what predictive model can we test our hypotheses or attempt to falsify our claims about that structure? Without such analytical constructs, linguistic research, whether Functional or Cognitive, must adopt an inductive, sample-based, empirical methodology. To these ends, experimental techniques for the analysis of semantics have been developed, yet usage-based methods remain poorly represented within the field.

Moreover, Cognitive Linguistics, a principal proponent of the Usage-based Model, recognises no internal language modules, such as syntax, lexis, or prosody. From a non-modular perspective, the study of meaning must account for the integration of all these components of language structure and do so simultaneously in a functionally and cognitively plausible manner. Corpus-driven methods, and multivariate statistics more specifically, are perfectly suited for such a task.

There are, therefore, two fundamental tasks ahead of us. Firstly, we must adopt inductive research methods. Whether elicited through experimentation,
extracted from electronic corpora, or collated from questionnaires and field research, generalisations based on samples of data present the only possibility for hypothesis testing, according to our model of language. Acknowledging this fact entails validating sample-based results through statistical confirmation. Secondly, we must develop corpus-driven semantic analysis. If we are to account holistically for the integrated complexity of the various dimensions of language structure, it is essential that we examine natural contextualised language production. Samples of natural language large enough to permit inductively valid claims are what we term corpora. Here again, statistics comes to the fore, though for a different reason. If we are to identify structure in this complex and dynamic system, multifactorial exploratory statistical analysis is a powerful, if not, essential tool.

The aim of this book is to both introduce quantitative corpus-driven semantic methodology to the broader research community and to advance the state-of-the-art. The methods in focus are those that are especially applicable to lexical and constructional semantic relations of similarity and difference. Linguistic forms are used in different ways, and capturing this semasiological variation is what we term the study of polysemy. Likewise, speakers choose between different linguistic forms to express similar concepts. Explaining this onomasiological variation is what we term the study of synonymy.

The reader should be aware that the term polysemy is not restricted to ‘true’ polysemy, where distinct referents are indicated by a single form. Instead, meaning is understood from a usage-based perspective, where any systematic variation in use represents semasiological structure. In the same vein, synonymy is not restricted to absolute similarity since, from a Cognitive Linguistic perspective, one assumes that any variation in form is motivated by some variation in use and that ‘true’ synonymy is rare, if it exists at all. Lastly, it must be added that the term semantics is used to indicate encyclopaedic semantics and pragmatics, as opposed to linguistic semantics in its narrow sense.

Moreover, the term corpus methodology should be understood to indicate research that is corpus-driven, as opposed to corpus-exemplified (q.v. Tummers et al. 2005). Where corpus-exemplified research identifies occurrences to explain or support a theory of language structure, corpus-driven research examines large samples of natural language in order to test theories about language structure (often previously proposed a priori in corpus-exemplified research). In this, we focus on quantitative techniques, and, more specifically, statistical methods for the exploration of data and the falsification or confirmation of quantitatively testable hypotheses.
We begin by developing operational definitions of polysemy and synonymy (section 2). The discussion then demonstrates, given the above definitions, why we need corpus methodology (section 3). Finally, we consider the argument that the quantitative corpus tradition and the prototype Cognitive Semantic tradition are not only analytically compatible, but, in fact, inherently entwined. It is argued that the introspection-based research in prototype structuring of linguistic categories and the results of the semantic network tradition should be understood as theoretical modelling of semantic structure, an integral first step in empirical research (section 4).

2. Polysemy and synonymy: Definition, object and operationalisation

Given the linguistic heritage of the 20th century, it is upon us to begin with two simple questions. Firstly, what exactly constitutes polysemous and synonymous semantic relations in usage-based semantics? Secondly, is it theoretically and analytically possible to speak of the polysemy and synonymy of a syntactic construction? This section answers each question in turn. In doing so, the discussion will offer operational definitions of the concepts ‘polysemy’ and ‘synonymy’ and justify extending the study of such relations in morpho-syntax.

With meaning defined as encyclopaedic, or simply ‘world knowledge’, and with all form being semantically motivated, how do we understand the notions of polysemy (difference in sense) and synonymy (similarity in sense)? With a traditional understanding of the terms, usage-based approaches, such as Cognitive Linguistics or Functional Linguistics, are not interested in polysemy or synonymy per se. Much of what cognitivists and functionalists examine would be called ‘semantic vagueness’, ‘word fields’ or ‘syntactic alternation’ in other theoretical paradigms. Moreover, for Structuralism, which holds a distinction between the langue and its use, our position, that the use of a word equates its meaning, is nonsense. For Generativism, which argues for an autonomous formal system, examining the semantic structure of syntactic patterns is equally nonsense.

In Structuralist terms, polysemy was identified using the definitional test and / or the ambiguity test. These tests were designed to distinguish polysemous relations from vague relations, on the one hand, and from a monosemic form, on the other. This modular understanding of semantic structure assumes two theoretical constructs - firstly, the notion of truth conditional semantics and, secondly, the notion of semantic categories determined by necessary and
sufficient conditions. The same assumptions determined the definition of synonymy – any two lexemes were considered synonymous if replacing one lexeme with the other, did not change the ‘truth semantic’ meaning of the phrase (Lyons 1968: 428). Cognitive Linguistics categorically refutes both assumptions. Indeed, many contemporary approaches to meaning have left behind the notion of necessary and sufficient conditions as well as any strict division between linguistic semantics and context pragmatics. Without such assumptions, the meaning of the terms polysemy and synonymy can be loosened and defined as:

Polysemy – conceptual-functional variation of a form
Synonymy – formal variation of a concept-function

Note that these definitions of polysemy and synonymy do not exclude taxonomic relations such as hyponymy and hyperonymy or basic level, superordinate and subordinate relations. The definitions need to include these relations because, in a given situation-context, a given form may be used at different levels of specificity (‘vertical’ polysemy), just as a choice is made between different words signifying different levels of specificity (‘vertical’ synonymy). Such a broad understanding of semantic relations could, perhaps, be more accurately described as semasiological and onomasiological variation (Geeraerts 1993b, Grondelaers & Geeraerts 2003). Nevertheless, we will continue with the terms polysemy and synonymy since they enjoy wider currency.

Can we now justify applying the notions of polysemy (semasiological variation) and synonymy (onomasiological variation) to the study of grammar? If we assume that all form is conceptually or functionally motivated and if we agree that in the study of polysemy and synonymy, we are, in fact, studying variation in concept-function relative to form and variation in form relative to concept-function, then this must necessarily be extended to non-lexical meaning and form. Therefore, in blunt terms, the study of polysemy and synonymy includes the study of schematic forms such as that typical of syntax and prosody just as much as it does the study of words and morphemes. We can, therefore, identify our object of study more precise terms:

Polysemy – the functional-conceptual variation of any symbolic form
Synonymy – the functional-conceptual relation between any symbolic forms

A simple statement about an object of study that many linguists would take as obvious, yet others take as ludicrous. The division is a result of the fact that
even if we accept that all formal structure is motivated, there certainly exist different types of form just as there exist different types of meaning. Phonological, gestural, syntactic, morphological and lexical forms all tend to possess different characteristics. There is no doubt in this – the referential meaning of a lexeme like *chair* is far from the intersubjective meaning of a request implicature, and these two ‘types’ of meaning differ from the abstract relational meaning of the Transitive Construction. Notwithstanding the belief that such linguistic devices are inextricably interwoven structurally, the characteristics of these different types of form and meaning, just as the tools needed to describe them, differ profoundly.

If we accept that, analytically (as opposed to theoretically), such differences exist, then we can distinguish different lines of research. Given this, it is still possible to speak of lexical research or syntactic research as long as this is seen as an analytical emphasis, not an object of study. In order to avoid conjuring the theoretical modules of other theories of language, we can term the different lines of research – schematic and non-schematic. Such distinctions are abstract enough to avoid leading us into the trap of modularising language structures, while also being transparent enough to help us easily differentiate between lines of research and the tools they necessitate.

Therefore, although semantic structure and its relations exist equally for morpho-syntax and lexis, the kind of semantics typically associated with lexical forms is of a much more concrete nature than that of more schematic formal structures. This is, of course, a tendency, but an important one because the different kinds of formal structure and the semantics associated with them may warrant different analytical techniques. For these practical purposes, let us identify four objects of study:

- **Polysemy**
  - Concrete meaning
  - Schematic meaning

- **Synonymy**
  - Concrete form
  - Schematic form

Seen in these terms, Lakoff’s (1987) analysis of *over* is a study of concrete, or non-schematic, polysemy, but his study of the Deictic Construction is schematic polysemy. On the other side of the coin, Lakoff’s (1987) analysis of *ANGER* is concrete (non-schematic) synonymy and the analysis of the Dative alternation by Goldberg (2002) is an instance of schematic synonymy. Regardless of whether we speak of words or syntax, the analytical object of the relations between different yet functionally-conceptually similar forms versus
the relations between different functions-concepts of a single form should be evident.

A few further examples should demonstrate the distinction and its value for identifying the fundamental objects of study for all usage-based linguistics. By way of example for schematic polysemy, consider Halliday’s (1967) study of the English Transitive Construction, Langacker’s (1982) analysis of the English Passive Construction or Bondarko’s (1983) analysis of the Perfective Aspect. For non-schematic polysemy, take Culioli’s (1990) French lexeme *donc* ‘so, thus’ or Fillmore’s (2000) lexeme *crawl*. For synonymy, the same diversity exists. At the schematic level, we have Givón’s (1982) evidential markers, Halliday’s (1985) English Grammatical Conjunctions, Talmy’s (1988) Causative Constructions, Culioli’s (1990) English Nominal Constructions. Obviously, the lexical research is equally diverse: from Fillmore’s (1977) BUY-SELL frame, the lexical field of SAY-SPEAK by Dirven et al. (1982), Lehrer’s (1982) study on adjectives for the description of wine to the full swath of conceptual metaphor and metonymy studies. Indeed, most descriptive linguistic studies can be classified as one of these four lines of research.

Despite the fact that a wide and varied range of linguistic analyses can be understood as the study of polysemy or synonymy, not all research can be characterised by this typology. Within both the Functional and Cognitive paradigms, there exist objects of study that are not readily characterised in this manner. Such research lies beyond the realm of polysemy and synonymy.

Within Cognitive Linguistics, *ad hoc*, or non-entrenched, categorisation typical of conceptual integration (Fauconnier & Turner 1998), just as the entire field of language processing are beyond the purview of polysemy and synonymy research. Within Functional Linguistics, the detailed, context dependent analysis of conversation or the wide-ranging research on the characteristics of genre, register, and stylistics seem to lie outside the realm of semantic relations *per se*. This is not to say that such research, both the cognitive and the functional, does not inform the study of semantic relations, but it is distinct from this object of study.

Having established the far-reaching importance and limits of these two objects of study – the conceptual-functional structure of the form and the forms available to express a concept-function, we can now ask how best to operationalise these notions? In other words, how can we define the object of study in a measurable way? With no *langue* or independent syntax against which we can test hypotheses, what exactly are we analysing in the study of semantic relations? How can we falsify or verify results? We need an opera-
tionalisation of the object of study that either offers stability to the system or a means of capturing the dynamic nature of that system. The answer lies in Langacker’s (1987: 59-60) theory of the entrenched form-meaning pair. The theory of entrenchment can be understood as an operationalisation of grammaticality – the more often a form-meaning pair is used, the more automated its processing becomes and the more ‘grammatically acceptable’ it is according to the speaker’s intuition. Generalised across a speech community, that is the ensemble of individual speakers’ linguistic knowledges, we have an operationalisation of grammar.

If we accept these hypotheses and assume that through repeated contextualised use, the relation between a concept-function and a form becomes stable, then we have an identifiable object of study. At a theoretical level, therefore, the study of semantic relations is the study of variation of entrenched form-meaning pairs:

- Polysemy – the entrenched functional-conceptual variation of a schematic or non-schematic form
- Synonymy – the entrenched functional-conceptual relation between schematic and non-schematic forms

It is important to note that Langacker’s theory of entrenchment is determined by frequency of use. This can be re-stated as an operational definition: the degree of entrenchment is determined by frequency of association of a given form and a given use of that form. Of course, what constitutes ‘a form’ and ‘a use’ is open to debate, but the notion of entrenchment per se is operationalised in a way that permits quantified analysis of semantic structure – the aim of this volume. The frequency of a form-meaning pair determines its entrenchment in a speaker’s knowledge. This, when extended to an entire speech community, means that the frequency of a form-meaning pair in language equals the degree of its stability in the intersubjective system of language. Therefore:

Polysemy and synonymy can be measured in terms of the relative frequency of association of form and meaning

Given this understanding of semantic relations, questions as diverse as prototype effects and sociolinguistic variation are neatly explained by this single principle. This claim deserves a brief explanation.

A central question for the study of polysemy is: which ‘senses’ are more ‘central’, or more prototypical, than others. A similar question exists in synonymy studies: which forms are more basic taxonomically (as in basic-level
terms, Lakoff 1987)? Both the concept of prototype meaning and basic-level form can be operationalised in terms of frequency. Although it is not claimed that frequency alone can explain prototype or taxonomic structure, it is, none-theless, one important operationalisation of these phenomena (see Arppe et al. 2010). The operationalised definition is straightforward: the more frequent a given meaning, the more ‘typical’ it is categorically. This is a frequency-based understanding of prototypicality. The same can be posited, mutatis mutandis, of basic level categories in taxonomic structure: if for polysemy, typicality is operationalised as the most frequent concept-function (relative to a form), for synonymy, basicness is operationalised as the most frequent form (relative to a concept-function). Therefore, for synonymy, the more frequent a form, the more basic it is taxonomically.

Importantly, this frequency-based understanding of the system integrates the varied and dynamic nature of language into our model of semantic relations. For a usage-based understanding of language, the system is emergent and entirely dependent on context – context of situation, context of speaker, context of time, context of region. A given form-meaning pair will be more frequent in one city than another, in one register than another, for one gender more than another, and at one period of time more than another. Relative frequency, at a theoretical level, eloquently incorporates this complexity into the object of study. Therefore, the operationalisation - context-sensitive frequency-based typicality – simultaneously captures semantic structure both categorically (prototype effects) and taxonomically (basic-level effects) but also relative to social variation.

Lastly, it must be stressed that this frequency-based approach to entrenchment is only an operational definition. Other operationalisations of the relationship between form and meaning may be equally valid. Despite Langacker’s concern with frequency, something that holds well for corpus linguists, perceptual and conceptual salience surely also have a hand in the learning process, and therefore entrenchment. Langacker’s explanation of entrenchment assumes that all input has the same ‘weight’ in or ‘impact’ upon the system. This, it would seem, is a simplification. There is no reason to suppose that every occurrence of, or exposure to, a language event has the same value in the process of entrenchment. The implications, especially for prototype and taxonomic structures are far reaching. We must suppose, therefore, that the frequency-based account of language cannot give us the full picture. It does, however, offer an operationalised and quantifiable object of study, one that will permit the testing of hypotheses, the verification of re-
sults, and clear benchmarks for the comparison of results, using other methodologies and other operationalisations.

3. Complexity and sampling: The need for quantitative techniques

Why are quantitative corpus methods needed for a cognitive approach to polysemy and synonymy? There are two answers to that question. The first answer takes us back to our object of study and the second, to our model of language.

The first reason we need quantitative techniques can be summarised as a question of complexity. We defined our object of study as “relative frequency of association of form and meaning”. The discussion has, thus far, ignored an important issue – what constitutes a given form and what constitutes a given meaning? We have stated that a form is any form and that meaning is anything we know of the world. In fact, both the terms ‘form’ and ‘meaning’ are entirely misleading. All forms exist in a formal context and are, therefore, composite. Theoretically, one should not speak of a form but of a composite form. Similarly, meaning is situated contextually and therefore one should not speak of a meaning as a reified sense, but as an intersubjective result of communication. The importance and implications of these two points cannot be underestimated.

Formal structure is complex. Even the simplest utterance is a composite form at some level. Moreover, dialect and sociolect variation means that even phonetic components can be indicative of usage variation. Prosody, syntax, morphology, lexis and even gesture, all come together in effectively every utterance as composite forms. Since it is a fundamental tenet of Cognitive Linguistics that language must be analysed holistically, we must treat forms as composite structures, always.

If formal structure is complex from a cognitive perspective, meaning is more so. Even lexical meaning cannot be divided into discrete senses. It is a dynamic, context dependent, multi-dimensional and intersubjective social phenomenon. Geeraerts’ (1993a) and Killgarriff’s (1997) studies on polysemy and word meanings mark important milestones in the study of semantic relations. Their work shows that lexical senses, just like any functionally or conceptually determined category, cannot be assumed to be discrete or reifiable. Arriving at this point, both theoretically and descriptively, was a long road. Via theoretical research on prototype structures, on the one hand, and via corpus research in lexicography, on the other, many researchers in lexical
Corpus method and cognitive theory

semantics have reached the conclusion that a reified and discrete understanding of semantic structure, including individuated senses, will not produce adequate descriptions of that structure.

Encyclopaedic semantics entails that all of sociolinguistic structure is included in the semantic analysis. The situation context, the gender, the age, the socio-economic class, the geographical region, and the social status are all dimensions of language use, dimensions of world knowledge in how to use language, how to communicate successfully. This world knowledge is encyclopaedic semantics and must therefore be integrated into semantic analysis. The result is a complex multidimensional ‘form’ coupled with a complex multidimensional ‘meaning’. It may be possible, using intuition and introspection to consider all of these dimensions, but to understand how they all interact, or how they could all interact, is an effectively impossible feat. Quantifying the analysis permits the use of multivariate statistics, which are designed for modelling and capturing structure in precisely this kind of complex system.

The second reason we need to develop quantitative methods for the study of semantic relations lies in the model of language propounded by Cognitive Linguistics. Both the Structuralist and Generativist traditions assumed models of language that permitted the falsification of claims made about its structure. Necessary and sufficient conditions, for establishing semantic categories, and grammatical acceptability tests, for checking proposed grammatical rules, both allow an analyst to falsify hypotheses. What possibility does Cognitive Linguistics have for falsifying propositions made about language structure? How can we test for the descriptive or explanatory adequacy of a conceptual metaphor or a reference–point construction? Neither makes any predictions that can be countered.

The lexical polysemy studies of early Cognitive Semantics were excellent examples of this problem. Lakoff (1987) proposed 21 senses for the lexeme over, but his work was challenged. Although this, it would seem, is good scientific procedure – a study proposes a given number of senses, the results are challenged – there is, ultimately, no way of resolving the issue because there is no way of disproving his original analysis. Table 1, below, lists the different proposals of the number of senses for over in English, Dutch and German. This debate could effectively continue ad nauseum, since using one’s intuition to determine a category, especially a category that can have both a fuzzy boundary and better or worse exemplars, has no possibility for falsification. It is, ultimately, a matter of opinion.
In previous approaches, truth conditional semantics and predictive rules could be tested using deductive proofs. Usage-based research has no such options. A counter example, even many, does not contradict a proposal about relative structure. From our perspective, linguistic structure is emergent, it is dynamic and varied. However, a quantification of the study of semantic relations permits inductive research. The only way to test our hypotheses is to take a sample of natural language use or a sample of elicited language use and make generalisations based on that sample. Once we speak in terms of samples and populations, then we speak of inductive research, and, in brief, statistics.

Table 1. Variation in polysemy analysis

<table>
<thead>
<tr>
<th>Lexeme</th>
<th>Senses Identified</th>
<th>Study</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>over</em> (English)</td>
<td>6 basic senses, 21 sub-senses</td>
<td>Lakoff (1987)</td>
</tr>
<tr>
<td></td>
<td>17 senses</td>
<td>Taylor (1989a)</td>
</tr>
<tr>
<td></td>
<td>9 of Lakoff’s sub-senses as 1 sense</td>
<td>Vandeloise (1990)</td>
</tr>
<tr>
<td></td>
<td>3 basic senses, &gt; 40 sub-senses</td>
<td>Deane (1993a, 2006)</td>
</tr>
<tr>
<td></td>
<td>6 basic senses, 12 senses sub-senses</td>
<td>Dewel (1994)</td>
</tr>
<tr>
<td></td>
<td>3 basic senses</td>
<td>Kreitzer (1997)</td>
</tr>
<tr>
<td></td>
<td>1 basic sense, 15 senses sub-senses</td>
<td>Tyler &amp; Evans (2003)</td>
</tr>
<tr>
<td><em>over</em> (Dutch)</td>
<td>13 senses</td>
<td>Cuyckens (1991)</td>
</tr>
<tr>
<td></td>
<td>3 basic senses, 11 sub-senses</td>
<td>Geeraerts (1992, 2006a)</td>
</tr>
<tr>
<td><em>über</em> (German)</td>
<td>3 basic senses, 8 sub-senses</td>
<td>Bellavia (1996)</td>
</tr>
<tr>
<td></td>
<td>3 basic senses, inconclusive sub-senses</td>
<td>Dewel (1996)</td>
</tr>
<tr>
<td></td>
<td>3 basic senses, 14 sub-senses</td>
<td>Meex (2001)</td>
</tr>
<tr>
<td></td>
<td>5 basic senses, 8 sub-senses</td>
<td>Liamkina (2007)</td>
</tr>
</tbody>
</table>

Statistical analysis gives us the probability that a given finding in a given sample is not chance, it gives the possibility of modelling the variation in our data and testing the accuracy of analyses by using these models to predict language use. Quite simply, moving towards quantification and statistical analysis appears inevitable for all usage-based language research.

4. **Modelling meaning. Multidimensional patterns and prototype effects**

How does the usage-based approach contribute to the prototype structure and semantic network tradition of polysemy and synonymy in Cognitive Semantics? The study of both (vague) polysemy and (near) synonymy has a long
tradition in Cognitive Linguistics. Indeed, many of the seminal works were devoted to such questions. Let us place quantitative corpus-driven research, such as that presented in the current volume and in Gries & Stefanowitsch (2006), Stefanowitsch & Gries (2006) and Glynn & Fischer (2010), in this ‘historical’ context.

At the end of the last century, what was often called the ‘network’ or ‘radial category’ approach to meaning included an wide range of polysemy and synonymy studies of especially spatial prepositions and grammatical cases. These forms were especially interesting because they linked the theoretical research on perception-based construal and image schemata with the culturally determined lexico-grammatical structure. The aim of this research was to model prototype structure and encyclopaedic semantics. In effect, presented with the boundless considerations of encyclopaedic semantics as well as relative categorisation due to prototype effects on structure, theoretically and analytically, much of the work can be seen as an attempt to identify order in what is an immensely varied and complex system. While theoretical models such as Frame Semantics (Fillmore 1985) and Idealised Cognitive Models (Lakoff 1987) attempted to explain the place of encyclopaedic semantics and prototype structuring in language, network analysis, in its various guises (Barthélemy 1991, Rice 1993, Geeraerts 1995), was essentially a representational formalism designed to visualise and summarise systematicity in semantic complexity. This formalism was used to various extents by different authors, but the principle of (i) employing encyclopaedic semantic features (ii) without the notion of necessary and sufficient conditions for category membership (iii) in order to distinguish senses and relate forms, underlies all semantic network research.


At first sight, this highly abstracted and introspective research tradition would seem distinct from, even at odds with, the bottom up approach of corpus-driven methodology. However, upon closer inspection, we see that the very origins of corpus-driven, indeed quantitative corpus-driven research, lie in the radial network studies. The contemporary methodology directly inherits from and builds upon this tradition. For both the study of synonymy and polysemy, many of the earliest studies were entirely empirical. Moreover, as we will see below, the two approaches are theoretically linked. Let us, however, begin with an overview of the radial network research.

For the practical concerns of brevity, it is impossible to offer even a snippet of this immense field. However, Table 2, below, offers a selection of studies, chosen to represent the depth and variation the radial network approach to semasiological structure. The object of study, its general part of speech or grammatical category, whether this is schematic or concrete in form as well as the method of analysis and reference are listed. It must be stressed that the distinction between schematic and concrete forms is only designed to show tendencies and no theoretical distinction is intended. The decision as to what to include is based on a subjective evaluation of the impact of the study and the author upon the field, as well the extent of the study, priority being given to monographs.

Research in synonymy, though it received less attention and possibly produced less in terms of quantity, was equally important in the development of Cognitive Semantics. Studies such as Fillmore (1977), Lehrer (1982), Dirven et al. (1982), Janda (1986), and later Schmid (1993), Geeraerts et al. (1994), and Rudzka-Ostyn (1995) represent seminal work in the field. Table 3, below, offers a summary of Cognitive Linguistic case studies in synonymy, again up until the turn of the century. There is some redundancy with table 2, because what may be a set of individual case studies on polysemy, were also combined to present a study in near-synonymy. Again, the table is designed to offer an overview and is no way complete.
Table 2. Prototype - encyclopaedic analysis of polysemy in Cognitive Semantics

<table>
<thead>
<tr>
<th>Object</th>
<th>Form</th>
<th>Schematicity</th>
<th>Method</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>lie (English)</td>
<td>verb</td>
<td>concrete</td>
<td>elicitation</td>
<td>Coleman &amp; Kay (1981)</td>
</tr>
<tr>
<td>talk, say, tell, speak (English)</td>
<td>verb</td>
<td>concrete</td>
<td>observation</td>
<td>Dirven et al. (1982)</td>
</tr>
<tr>
<td>over (English)</td>
<td>prep.</td>
<td>concrete</td>
<td>introspection</td>
<td>Brugman (1983a)</td>
</tr>
<tr>
<td>on the go Cx. (English)</td>
<td>constr.</td>
<td>schematic</td>
<td>introspection</td>
<td>Brugman (1983b)</td>
</tr>
<tr>
<td>up, out (English)</td>
<td>particle</td>
<td>concrete</td>
<td>introspection</td>
<td>Lindner (1983)</td>
</tr>
<tr>
<td>idea (English)</td>
<td>noun</td>
<td>concrete</td>
<td>introspection</td>
<td>Brugman (1984)</td>
</tr>
<tr>
<td>kind of Cx. (English)</td>
<td>constr.</td>
<td>schematic</td>
<td>introspection</td>
<td>Kay (1984)</td>
</tr>
<tr>
<td>out (Dutch), wy (Polish)</td>
<td>prep.</td>
<td>concrete</td>
<td>introspection</td>
<td>Rudzka-Ostyn (1985)</td>
</tr>
<tr>
<td>spatial preps. (Eng.)</td>
<td>prep.</td>
<td>concrete</td>
<td>introspection</td>
<td>Herskovits (1986, 1988)</td>
</tr>
<tr>
<td>zur, pere-, do-, at- (Russian)</td>
<td>prefix</td>
<td>concrete</td>
<td>introspection</td>
<td>Janda (1986)</td>
</tr>
<tr>
<td>spatial preps. (French)</td>
<td>prep.</td>
<td>concrete</td>
<td>introspection</td>
<td>Vandeloise (1986)</td>
</tr>
<tr>
<td>over (English)</td>
<td>prep.</td>
<td>concrete</td>
<td>introspection</td>
<td>Lakoff (1987)</td>
</tr>
<tr>
<td>let alone Cx. (English)</td>
<td>constr.</td>
<td>schematic</td>
<td>introspection</td>
<td>Fillmore et al. (1988)</td>
</tr>
<tr>
<td>down (English)</td>
<td>prep.</td>
<td>concrete</td>
<td>introspection</td>
<td>Schulze (1988)</td>
</tr>
<tr>
<td>tall (English)</td>
<td>adj.</td>
<td>schematic</td>
<td>elicitation</td>
<td>Dirven &amp; Taylor (1988)</td>
</tr>
<tr>
<td>ask (English)</td>
<td>verb</td>
<td>concrete</td>
<td>observation</td>
<td>Rudzka-Ostyn (1989)</td>
</tr>
<tr>
<td>Genitive (English)</td>
<td>case</td>
<td>schematic</td>
<td>introspection</td>
<td>Taylor (1989b)</td>
</tr>
<tr>
<td>Dative (Czech)</td>
<td>case</td>
<td>schematic</td>
<td>introspection</td>
<td>Janda (1990, 1993)</td>
</tr>
<tr>
<td>vers (Dutch)</td>
<td>adj.</td>
<td>concrete</td>
<td>introspection</td>
<td>Geeraerts (1990)</td>
</tr>
<tr>
<td>Middle Voice Cx. (French)</td>
<td>constr.</td>
<td>schematic</td>
<td>observation</td>
<td>Melis (1990)</td>
</tr>
<tr>
<td>Resultative Cx. (English)</td>
<td>constr.</td>
<td>schematic</td>
<td>introspection</td>
<td>Goldberg (1991, 1995)</td>
</tr>
<tr>
<td>Dative (Polish)</td>
<td>case</td>
<td>schematic</td>
<td>introspection</td>
<td>Rudzka-Ostyn (1992, 1996)</td>
</tr>
<tr>
<td>over (Dutch)</td>
<td>prep.</td>
<td>concrete</td>
<td>introspection</td>
<td>Geeraerts (1992)</td>
</tr>
<tr>
<td>Ditransitive Cx. (English)</td>
<td>constr.</td>
<td>schematic</td>
<td>introspection</td>
<td>Goldberg (1992, 1995)</td>
</tr>
<tr>
<td>in (Dutch)</td>
<td>prep.</td>
<td>concrete</td>
<td>introspection</td>
<td>Cuyckens (1993)</td>
</tr>
<tr>
<td>Instrumental (Russian)</td>
<td>case</td>
<td>schematic</td>
<td>introspection</td>
<td>Janda (1993)</td>
</tr>
<tr>
<td>at, on, in (English)</td>
<td>prep.</td>
<td>concrete</td>
<td>introspection</td>
<td>Rice (1993)</td>
</tr>
<tr>
<td>at, by, to (English)</td>
<td>prep.</td>
<td>concrete</td>
<td>introspection</td>
<td>Newman (1993)</td>
</tr>
<tr>
<td>(a)round (English)</td>
<td>prep.</td>
<td>concrete</td>
<td>introspection</td>
<td>Deane (1993b)</td>
</tr>
<tr>
<td>Genitive (Polish)</td>
<td>case</td>
<td>schematic</td>
<td>introspection</td>
<td>Schulze (1993)</td>
</tr>
<tr>
<td>in (English)</td>
<td>prep.</td>
<td>concrete</td>
<td>introspection</td>
<td>Rudzka-Ostyn (1994)</td>
</tr>
<tr>
<td>Instrumental (Polish)</td>
<td>case</td>
<td>schematic</td>
<td>introspection</td>
<td>Vandeloise (1994)</td>
</tr>
<tr>
<td>off (English)</td>
<td>prep.</td>
<td>concrete</td>
<td>introspection</td>
<td>Dąbrowska (1994)</td>
</tr>
<tr>
<td>op (Dutch)</td>
<td>prep.</td>
<td>concrete</td>
<td>introspection</td>
<td>Cuyckens (1994)</td>
</tr>
<tr>
<td>over (English)</td>
<td>prep.</td>
<td>concrete</td>
<td>introspection</td>
<td>Dewel (1994)</td>
</tr>
<tr>
<td>answer, respond (English)</td>
<td>verb</td>
<td>concrete</td>
<td>observation</td>
<td>Rudzka-Ostyn (1995)</td>
</tr>
<tr>
<td>door, langs (Dutch)</td>
<td>prep.</td>
<td>concrete</td>
<td>introspection</td>
<td>Cuyckens (1995)</td>
</tr>
<tr>
<td>Caused-Motion Cx. (English)</td>
<td>constr.</td>
<td>schematic</td>
<td>introspection</td>
<td>Goldberg (1995)</td>
</tr>
<tr>
<td>at, on, in (English)</td>
<td>prep.</td>
<td>concrete</td>
<td>elicitation</td>
<td>Sandra &amp; Rice (1995)</td>
</tr>
<tr>
<td>Dative (Polish)</td>
<td>case</td>
<td>schematic</td>
<td>introspection</td>
<td>Dąbrowska (1997)</td>
</tr>
<tr>
<td>over (English)</td>
<td>prep.</td>
<td>concrete</td>
<td>introspection</td>
<td>Kreitzer (1997)</td>
</tr>
<tr>
<td>figure out (English)</td>
<td>verb</td>
<td>concrete</td>
<td>introspection</td>
<td>Morgan (1997)</td>
</tr>
<tr>
<td>at, on, in (English)</td>
<td>prep.</td>
<td>concrete</td>
<td>elicitation</td>
<td>Cuyckens et al. (1997)</td>
</tr>
<tr>
<td>Causative Cx. (English)</td>
<td>constr.</td>
<td>schematic</td>
<td>observation</td>
<td>Lemmens (1998)</td>
</tr>
<tr>
<td>Dative (Dutch)</td>
<td>constr.</td>
<td>schematic</td>
<td>introspection</td>
<td>Geeraerts (1998)</td>
</tr>
<tr>
<td>straight (English)</td>
<td>adj.</td>
<td>concrete</td>
<td>introspection</td>
<td>Cienki (1998)</td>
</tr>
<tr>
<td>on (English)</td>
<td>prep.</td>
<td>concrete</td>
<td>observation</td>
<td>Rice et al. (1999)</td>
</tr>
<tr>
<td>to, for (English)</td>
<td>prep.</td>
<td>concrete</td>
<td>observation</td>
<td>Rice (1999)</td>
</tr>
<tr>
<td>What’s X doing Y Cx. (Engl.)</td>
<td>constr.</td>
<td>schematic</td>
<td>introspection</td>
<td>Kay &amp; Fillmore (1999)</td>
</tr>
<tr>
<td>crawl (English)</td>
<td>verb</td>
<td>concrete</td>
<td>introspection</td>
<td>Fillmore (2000)</td>
</tr>
</tbody>
</table>
It should be noted that although there was less work on synonymy per se, the conceptual metaphor studies, which were extremely numerous at the time, are, in effect, synonymy studies. Although such research was primarily interested in figurative lexemes, they remain, nonetheless, studies of near-synonymy (as noted by Kittay & Lehrer 1981 at the time). There was much discussion about what constitutes a source domain and / or target domain and whether certain expressions were in fact examples of the concept in question. From a lexical point of view, these questions are, of course, questions of near-synonymy. Expressions such as to have the hots versus to be head over heels were said to profile different aspects of the target domain, in other words, they were near-synonyms.

In order to appreciate the trends and heritage of contemporary methods in Cognitive Semantic research, it is helpful to make a quantified summary of the research output. Since, for reasons of space, full coverage of this research history is impossible, only the results of the investigation are presented. A
survey of 126 studies was conducted from roughly the beginning of the Cognitive Linguistic research community with the publication of Paprotté & Dirven’s (1985) anthology *The Ubiquity of Metaphor* to the turn of the century, where the field becomes extremely diverse and empirical methods begin to become the norm.

In order to operationalise the criterion of what constitutes ‘cognitive’ research in the field, the survey is restricted to three publication avenues. Only full articles that addressed polysemy or synonymy in the three following kinds of sources are considered:


(ii) Five foundational anthologies within the paradigm, three of which are proceedings of the first three conferences of the International Cognitive Linguistics Society (Rudzka-Ostyn & Geiger 1993, de Stadler & Eyrich 1993, and Casad 1996) and two of which predate the society but, at the time, were constitutive of the community (Paprotté & Dirven 1985 and Rudzka-Ostyn 1988).


Monographs are not included since it is difficult to gauge the impact and importance they contributed to the field but also because they often contain multiple case studies. Another issue is how to determine what constitutes an example of the network approach to semantic relations. This is determined using the three-part definition offered above: employing encyclopaedic semantic features; without the notion of necessary and sufficient conditions for category membership; in order to distinguish senses or relate forms.

Each of the 126 studies are categorised for year of publication, type of publication and author. The object of study for each study is also categorised as:

- schematic vs. concrete (lexemes vs. constructions / grammatical categories)
- polysemy vs. synonymy
- linguistic phenomenon (actual form(s) under investigation)
- language
Lastly, the studies are categorised for their method of analysis. Three kinds of method are distinguished: introspection, observation, and elicitation. Two of these methods are further distinguished. For observational data (corpus based) methods, three methods are distinguished:

- corpus-driven with statistical verification
- corpus-driven with raw counts
- corpus-illustrated (introspection exemplified with natural data)

For elicited methods, another 3 methods are distinguished:

- quantified direct elicitation (questionnaires etc.) with raw counts
- quantified direct elicitation with statistical verification
- experimental elicitation with statistical verification

Distinguishing introspective, experimental and observational methods is unproblematic, save when a study uses more than one method, as in Dirven & Taylor (1988). In this case, for instance, the study is categorised as elicited, since these data feature more prominently than the corpus data in the analysis.

The most striking result is just how balanced and broad the range of studies is. Although obviously Eurocentric, a surprisingly wide range of languages is considered (in total 30 different languages, with English making up 36% of the studies). Moreover, despite the predilection for spatial prepositions, particles, and morphemes, these parts of speech represent ‘only’ 18 (14%) of the studies. Although 14% represents a sizable proportion of the research, considering that the approach is often termed preposition research and considering that two of the anthologies were devoted to prepositions and another two to spatial representation, this figure is not as overwhelming as one might expect.

However, the best indicator of the diversity of the research is found if we take a more coarse-grained perspective and consider schematic vs. concrete and synonymous and vs. polysemous objects of study. Figure 1, below, represents the numbers of such studies over the 15-year period.

We see here how evenly dispersed the four different objects of study are over the period. Only in 1993, do we see a divergence, where the number of studies examining schematic forms such as grammatical constructions and grammatical categories drop when other objects of study remain steady or increase in number. Indeed, divided in this manner, not even the analysis of
concrete instances of polysemy, such as prepositions, is even vaguely dominant.

**Figure 1.** Object of study in Cognitive Semantics 1985-1999

Although this tells us nothing of the methodological heritage, it is important to note that semantic network analysis was not restricted to polysemy studies of spatial prepositions, and that the diverse range of lexical and grammatical analysis of both near-synonymy and vague-polysemy visible today is directly descendent from this tradition.

Turning to the methodological trends, we see an important and consistent presence of empirical studies, even if the use of introspection dominates. Figure 2 depicts these trends. Two levels of granularity are summarised in the single plot: a simple distinction between empirical and introspective as well as a break down of empirical into elicited and observational.

Figure 2, below, clearly represents how empirical methods followed the trends in introspective studies and were, although less common, far from irrelevant. The exception to this is, again, 1993, where we see a large number of introspective studies and no published observation or elicitation-based research in the sample.
Figure 2. Method of study in Cognitive Semantics, 1985-1999

At a more fine-grain level, there is a striking lack of statistical sophistication in the observation-based research. In the sample of studies, only a single corpus analysis, Gries (1999) showed any sophistication, the entire body of corpus-driven research being restricted to raw counts. Although, we know this was not completely the case since studies outside the narrow range under consideration employed statistical techniques from the beginning of the research paradigm, the relative trend is clear – the use of statistical techniques to consider results of corpus-driven research was not the norm. The use of statistical techniques for elicited data, however, was more common, though far from typical and appearing quite late in the sample, Schulze (1991), Chaffin (1992) and Myers (1994) being the earliest instances.

Although based on a limited sample of 126 studies from the official journal and a selection of anthologies, the survey hopefully demonstrates that the semantic tradition within Cognitive Linguistics is not restricted to prepositional studies nor is its methodology overwhelmingly introspective. Perhaps the image that Cognitive Linguistics is exclusively orientated towards introspective methodology results from the most widely known theoretical works that founded the paradigm. Certainly the seminal publications of Fillmore (1985), Talmy (1985), Lakoff (1987), and Langacker (1987), restrict themselves to introspective investigation. The importance of these theoretical and foundational works notwithstanding, we see above that the research commu-
Corpus method and cognitive theory

Corpus method and cognitive theory as a whole, always included significant methodological diversity. Quantitative corpus-driven methods are therefore, not a new turn or a new direction, but the natural development of an existing tradition.

The rich descriptive heritage summarised in tables 2 and 3 represents the Cognitive Semantic approach to polysemy and synonymy research. This volume continues the tradition, but instead of turning to prototype set theory as the analytical framework to capture the complexity of semantic relations, the corpus-driven research presented here uses multivariate statistical modelling and collocation association measures. Does this mean that we have done away with lexical network analysis, prototype category structure and fuzzy set structure? It would be possible to dismiss the introspective semantic network analyses of the early period of Cognitive Linguistics as little more than exercises in prototype set theory. After all, it has been shown that the method of analysis was largely ad hoc (Sandra & Rice 1995) and the last decade has seen few important studies employing the approach. Indeed, many in the empirical research community would seek to distance themselves from such a research tradition. However, to turn our back on this tradition or seek to demonstrate the superiority of the current methodology would do injustice to this previous research.

Firstly, it is precisely this research tradition that freed the study of semantic relations from the notions of discrete senses and context independent semantics. Lexical network studies were the first and essential step towards this realisation – both theoretically and analytically. Theoretically, it set the stage for an understanding of meaning as emergent structure and, analytically, it produced the idea of understanding meaning as a network of interacting factors. On both fronts, the corpus-driven and experimental study of semantic relations is a direct heir to lexical network research. Indeed, within the tradition, as early as Geeraerts (1993a: 260), the idea that we need to move away from a reified and mono-dimensional understanding of meaning was being overtly mooted.10

Secondly, such studies are an essential step in empirical research. They represent hypothetical models of language structure, based on careful and systematic introspection-based analysis of language. Rather than ignore, or worse still, dismiss such theoretical research, empirical analysis needs to treat it as foundational. The results of introspection-based research are theoretical models, models that are, most likely, accurate descriptions of language structure. The new generation of experimental and observational linguistic analysis needs to test the accuracy and explanatory power of those models, modifying them where needed. Seen in this light, the configuration of the argument-
structure of the verbs of ‘buying and selling’ by Fillmore (1977), the schema specifications in Lakoff’s (1987) study of over, or the image-schematic grammatical features of Janda’s (1993) study of the Dative are theoretical models and it is our task to test their accuracy and improve them.

There are, of course, differences between the corpus-driven quantitative research and the introspection-based lexical network studies. No matter how large a corpus, found data will always be biased towards what is common rather than what is possible. Introspection is a vital methodology for proposing hypotheses about what is possible in a language. It follows that a lack of corpus evidence of a given form-meaning pair, does not mean it is not possible or does not occur. This is simply because even the largest corpus in the world is but a microscopic fraction of actual language use.

This difference affects the results profoundly. Corpus-driven research is exclusively frequency-based, and this, in turn, will exclusively reveal typical structures over less typical structures. It is for this reason that it would be difficult to simply carry out a corpus-driven study on over and compare the results with Lakoff’s (1987) results. A corpus-driven study may or may not confirm parts of the network analysis, but it surely would not paint the same picture anymore than lack of confirmation would negate his hypothesis. This is simply because only the most frequent usage patterns, or schema configurations to use Lakoff’s terminology, would be found.

Therefore, we see that, despite important differences, there is a direct line of descent in the methodology from the semantic network approach described above to the contemporary corpus-driven research such as Gries (2006), Divjak & Gries (2006), Divjak (2006, 2010), Wulff (2006), Janda & Solovyev (2009), Glynn (2009, 2010a, 2010b, forthc.) and Speelman & Geeraerts (2010). But what has become of the analytical apparatus - prototype categorisation and fuzzy sets? The results of multifactorial analysis and collocation analysis are, in fact, structured as non-reified prototype categories. Although there may be little overt reference to the prototype effects revealed through multifactorial or collostructional analysis, the results are based upon relative frequency and are, therefore, necessarily ‘prototype’ structured (at least if we accept a frequency-based operationalisation of prototypicality). Moreover, the multifactorial feature analysis identifies ‘meanings’ as tendencies, where a tendency is a multidimensional pattern of use. This, quite literally, produces networks of different uses – a frequency-based and complex multidimensional network of sense relations. The lexical network analysis produced prototype maps of meaning upon one ‘semantic’ dimension where the ‘nodes’ were discrete reified senses. Multifactorial feature analysis produces multidimen-
Corpus method and cognitive theory

Corpus method and cognitive theory

Notes


2. This definition would also cover antonymy. Although it may seem a little far-fetched to consider antonymy as a synonymy relation, this is merely a result of the terminology. Ideally, onomasiological structure would be a better term than synonymy, but the term is not established in the Anglo-Saxon tradition and would, therefore, add considerable terminological weight to the discussion. Nevertheless, it must be noted that antonymy should not be seen as an antonym (in a non-technical sense) for synonymy. It is well know from the study of lexical fields and, more recently, word space modelling in computational linguistics, that antonyms are, in fact, closely related to each other semantically and, therefore, relatively synonymous. For example, hate is much closer in meaning to love than car or run. Generally, antonyms are synonyms semantically opposed by one culturally or perceptually salient feature. Therefore, in fact, antonymy is an antonym to synonymy, in other words, very close in its meaning. See Jones (2002: 51) and Murphy (2003: 37) for examples of the issue at hand.

3. Seen from this perspective, the entire tradition of conceptual analysis (Wierzbicka 1985, Lakoff 1987, Stepanov 1997, Borkachev 2004, and Bartmiński 2008) is based upon synonymy. Such research begins with a concept and examines what words or expressions are available for its linguistic representation. On the grammatical front, a similar notion holds. Understood in these terms, Talmy’s (1988) Force Dynamics or Langacker’s Causative Constructions (1991: 408-411) are essentially onomasiological fields of near-synonymous forms. Of course, regardless of whether it is lexis or syntax, in order to understand the relationship between these
different forms, we must investigate each form semasiologically, that is, its polysemy. Bondarko (1991) argues convincingly that the semasiological - onomasiological divide is fundamental to any theory of conceptually or functionally motivated language.

4. For a discussion on the importance of operationalisation in language science, see Stefanowitsch (2010).

5. See also Givón’s (2005: 48ff) more detailed investigation into what he terms automated processing in the attention system. The idea of form-function mapping is extended to include perceptual and conceptual issues such as prototype structure.

6. The theory of entrenchment is based on the notion of automatisation, a widely accepted theory in psychology. Schneider & Shiffrin (1977) and Shiffrin & Schneider (1977) were the first to develop the hypothesis.

7. It is not the point of this discussion to enter into the debates on sign and communication theory. Suffice it is to say that any sign theory to which a Cognitive Linguist would ascribe, would also see meaning as a result of a communicative act and not an inherent objectifiable phenomenon.


10. Indeed, this line of thinking is not as radical as one might expect. Lehrer & Lehrer (1994) and Victorri & Fuchs (1996) represent examples of early discussions on how a non-reified understanding of semantic structure needs to be developed.

References


Dylan Glynn 229


Gries, Stefan Th. 2006. Corpus-based methods and Cognitive Semantics: The many senses of to run. In St. Th. Gries & A. Stefanowitsch (eds.), Corpora in Cogni-
Dylan Glynn


Lindner, Susan. 1983. *A Lexico-Semantic Analysis of English Verb-Particle Constructions with up and out*. Trier: LAUT.


Corpus method and cognitive theory


Sandra, Dominiek & Sally Rice 1995 Network analyses of prepositional meaning: Mirroring whose mind – the linguist’s or the language user’s? Cognitive Linguistics 6: 89–130.


