Synonymy, Lexical Fields, and Grammatical Constructions. A study in usage-based Cognitive Semantics

Dylan Glynn

1. Introduction

Cognitive Linguistics is, by definition, a usage-based approach to language. Its model of language places usage at the very foundations of linguistic structure with a linguistic sign, the form-meaning pair, argued to become entrenched through repeated successful use. It is this entrenchment that renders symbolic gestures linguistic rather than merely incidental and represents the key to structure in language. Patterns of language usage across many individuals can be argued to be indices of shared entrenchment. When large numbers of language users possess the same or similar entrenchment, we can talk about grammar, that is, linguistic structure.

Importantly, as cognitive linguists, we believe this structure to be conceptually motivated. A basic phenomenon in conceptual structuring is salience. This concerns the conceptual prominence of perceived (or conceived) objects and their relations. Although frequency represents an important factor in determining salience, a one-to-one relationship between relative frequency and relative salience does not exist. Various cultural and perceptual factors can make relatively infrequent concepts salient and vice versa. Corpus-driven linguistics is frequency based and so inherently restricted in what it can say about conceptual salience. Nevertheless, frequency data are perfectly placed to allow us to make generalisations about patterns of usage across speech communities. Importantly, from a Cognitive Linguistics perspective, we can make the assumption that these patterns of usage represent speakers' knowledge of their language, including the conceptual structures that motivate language. In this indirect way, the inductive generalisations based on frequency permit us to make hypotheses about the conceptual structure of language. This is possible without making more theoretically tenuous claims about the relation of frequency to cognition, such as those presented in Gries (1999) and Schmid (2000).[1]

Glynn, D. 2010. Synonymy, Lexical Fields, and Grammatical Constructions. A study in usage-based Cognitive Semantics. In *Cognitive Foundations of Linguistic Usage-Patterns*, H.-J. Schmid & S. Handl (eds), 89-118. Berlin: Mouton de Gruyter.

This study examines new usage-based techniques to capture semantic relations between near-synonymous words. The conceptual space encoded by a language is divided up in complex ways by lexical semantics. It follows that the study of lexical synonymy has a long tradition within Cognitive Linguistics. Moreover, the tradition dates back to some of the first corpus-driven research within the cognitive framework. Beginning with Dirven et al. (1982), Lehrer (1982), Schmid (1993), Geeraerts, Grondelaers and Bakema (1994), and Rudzka-Ostyn (1995) a strong line of empirical research developed. The current state of the art divides into the study of lexical near-synonyms (Newman & Rice 2004a, 2004b, Divjak 2006, Divjak & Gries 2006) and syntactic alternations (Gries 1999, Heylen 2005, Grondelaers et al. 2007, Speelman & Geeraerts forthc.).[2] This study advances upon previous approaches by applying a different statistical technique and by experimenting with direct semantic analysis in the annotation.

Within Cognitive Linguistics, the use of corpora and empirical methods more generally represents an important movement. Indeed, many argue that such approaches are crucial to the advancement of the field (Geeraerts 2006, Gibbs 2007, Croft 2008). The application of such methods to the study of semantics is not, however, straightforward. Corpus linguistics is essentially the analysis of large numbers of examples. A corpus linguist must examine many hundreds or even thousands of utterances before he or she can make any generalisations. It must be remembered that those generalisations are only valid to the extent that the analysis of those examples is valid. It is a common myth that corpus linguistics replaces linguistic analysis with quantitative deductions. Nothing is further from the truth. The annotation of a dataset is the laborious linguistic analysis of examples. Often computational techniques allow one to automate much of that analysis, but in the field of semantics, this is not possible. This study is concerned with precisely these quantitative usage-based methods for semantic description and so annotation is entirely made up of manual semantic analysis.

2. BOTHER: Lexical Field, Conceptual Space, Three Near-Synonyms

2.1. Near-Synonymy and Grammatical Constructions

Synonymy, or more precisely near-synonymy, is the study of semantic relations between lexemes or constructions that possess a similar usage. In this study, we focus on three lexemes denoting the concept BOTHER; these

are *annoy*, *bother*, and *hassle*. Example (1) captures the kind of semantic relations in question. We seek to explain speaker choice between these lexemes.

(1) People need paypal.... Too much *hassle* over cheques, especialy when you cant be *bothered* to check your statement, god she *annnoyed* me.[2]

Closely related lexemes have a special place in Cognitive Linguistics because their use, both in terms of their overlap and difference, can be seen as a reflection of the conceptual structures that motivate language use, and thus its structure. Although there is a certain circularity in this reasoning, we can justify approaching the question in such terms because speakers choose between linguistic forms when they speak. If we assume that speakers have knowledge of their language and culture and make their judgements based on that knowledge, this entails that their choices will reflect such knowledge. In Cognitive Linguistics, where entrenched language structure (or knowledge of language use) equates conceptual structure, by identifying the patterns of similar and distinctive usage, we chart the conceptual structure that motivates those patterns.

The principle is the same for the study of polysemy. Indeed, the cognitive study of polysemy and near-synonymy can be seen as a re-working of the Structuralist semasiological - onomasiological distinction (see Geeraerts, Grondelaers and Bakema 1994). Seen in this light, polysemy, or semasiological variation, is the study of the different uses of a form and synonymy, or onomasiological variation, is the study of the choice between different forms. If we make generalisations about usage based on large numbers of examples, then we have a usage-based approach to conceptual structure. This, of course, must be presented with the caveat that we cannot make clear deductions about conceptual categorisation and prototypicality until the relationship between ontological salience and frequency of use is better understood.

However, it is too simplistic to speak of choices between words. Just as lexical choices are reflections of different construals, so too are their grammatical expression. The belief that different 'lexicogrammatical framings' or 'configurational structurings' that result from the integration of lexical semantics and different parts or speech and morpho-syntactic forms represents a fundamental tenet of Cognitive Linguistics (Fillmore 1977: 128, Langacker 1987: 138ff, Talmy 1988: 173ff, Fillmore 2003: 250f).

When a speaker wishes to express the concept of BOTHER, for instance, it is unlikely that the speaker decides beforehand and independently of the context that this concept will be profiled nominally or verbally, just as it is unlikely that, given a verbal choice, he or she will have a predetermined selection between encoding the concept as an intransitive or transitive event. The ability to construe events and things, of even the most concrete nature, means that it will be rare that the speaker has no choice in this matter. If we can assume that the kinds of grammatical semantics associated with grammatical class and grammatical construction are part of the semantics expressed by the speaker, then they are an integral part of the lexeme chosen. It is for this reason that we cannot consider only verbs or only nouns in the study of synonymy.

There are two points to consider here. Firstly, grammatical semantics are not predictable "additions" to the lexical semantics. Although often the grammatical profiling of a lexical concept results in regular semantic integration, that is not always the case (Glynn 2002, 2005, 2007, forthc.). Therefore, we need to treat the interaction between the different grammatical profilings of the lexical concept as onomasiological choices, that is, part of the synonymous field. Secondly, there is growing evidence that language knowledge is largely redundant and that speakers rote-learn large amounts of profiling variation as entrenched units (Dąbrowska 2006). This means, for example, the simple and the continuous form of a verb or the nominative and instrumental case of a noun are entrenched as separate linguistic units and not 'generated by the grammar'. This is in line with Croft's (2001) arguments for a fundamental Construction Grammar approach to language structure. For these two reasons, the semantic unpredictability of lexical-grammatical composition and the fact that many of these compositions are entrenched as separate form-meanings pairs, if we are to produce a cognitively realistic grammar of lexical choice, we cannot restrict ourselves to one part of speech. Since from a Construction Grammar point of view, parts of speech are merely a subtype of grammatical constructions, we will refer to this formal variation as *grammatical class* and assume there is only a theoretical divide between the formal variation of grammatical class and grammatical construction.[3]

There is one last complication that must be taken on board in a usage-based approach to synonymy. Since generalising about the entrenched usage of many individuals is the basis of our grammar, we must account for variation between those individuals and within that usage. Therefore, Cognitive Semantic study, as a usage-based approach, must necessarily include

what is traditionally considered extralinguistic and social parameters, such as register and dialect.[4] By including this information, we achieve a truly usage-based description of usage patterns relative to a range of factors such as age, sex, region, language mode, and register.

We can conclude that the study of lexical near-synonymy is important and informative from a Cognitive Linguistic perspective since it offers us an indirect method for mapping conceptual structure via lexical choices. However, these lexical choices interact in a complex way with formal variation and the grammatical profiling of those lexical concepts. We need, therefore, to treat near-synonymy across the various grammatical classes and grammatical constructions that combine with lexical concepts. Lastly, choice between these forms is made in the social context of their use. Variation between language users and speech contexts surely affects lexical choice and so these dimensions must also be added to the equation.

We are, therefore, confronted with an inherently multidimensional object of study. We must identify patterns in usage relative to a wide range of forms and relative to a wide range of contexts. It is this multidimensional element of language structure that calls for the use of multifactorial statistical techniques to help identify usage patterns. This aspect of usage is not so readily accessible employing intuitive methods of analysis. Indeed, the multidimensional element of language structure is not identifiable when one considers the frequency of the different factors of usage individually. We need to access the simultaneous interaction of the different factors of language and to do so we need multifactorial techniques. This study demonstrates why such an approach is necessary and considers one simple technique for its application. In contrast with previous quantitative studies of synonymy, which have employed Hierarchical Cluster Analysis (Divjak 2006, Divjak & Gries 2006), we employ a technique not previously used for such purposes. This technique, Correspondence Analysis, has the advantage that it maps correlations rather than simply grouping variables. It has, however, the disadvantage that its visualisations can be difficult to interpret.

2.2. Data and Analysis

The data for this study comes from a large non-commercial corpus built from on-line personal diaries. The language is informal and in many ways similar to spoken mode. In part, this is due to the "Dear Diary" writing tradition that involves talking 'to your diary', but it is also because these

diaries 'speak back'; the LiveJournal on-line diary service used to build the corpus is interactional. This service allows the readers to respond to the "blog" entries and they regularly do. Indeed, the authors expect it and they often complain when their readers do not enter into dialogue. The corpus is made up of diary entries proper, not the dialogues, but the monologic-dialogic distinction is blurred since the writer is assuming that people will respond to his or her text. Evidence of this may be found in the countless references to certain readers and frequent switching to second person, both singular and plural. This results in quite a unique discourse style that is at once narrative and dialogic.

Despite the richness of the language in its naturalness, the corpus represents only a single text type. This is a basic and inherent limitation for this study. Corpus representativity is an important and often under estimated issue for usage-based approaches to language. One must be careful not to draw conclusions about language based on a single corpus, but at most about the language type represented in that corpus. For our purposes, the fact that we consider lexemes that differ in register but we have only one text type, which is of a most informal nature, is a serious shortcoming. However, one of the advantages of corpus driven research is that a study may be repeated on a second corpus and the results compared. For the current purposes, which are to demonstrate the viability and usefulness of the method, the on-line diary corpus suffices. Needless to say, further research will be necessary to confirm the results. This is true for both the need of confirmatory statistical analysis as well as verification through repeat analysis on different data.

From this corpus a relatively even number of the three lexemes were extracted, each with considerable context, totalling approximately 2,000 observations. Across these examples, the proportion of the different parts of speech, or grammatical classes, for each lexeme is maintained as it occurs in the corpus. The kind of formal variation in question is best described by way of example. Examples (2a) - (2h) summarise each of the major class-construction formal variants in question and serve to introduce the kind of language that is typical of the corpus.

(2) a. Saw quite a few people I knew, including the awful stalker guy who's been hassling me ... (Transitive)

b. hassle me, bother me, bug me, give me a bad time, If you hassle me about my kinky hair, I'll cut it all off. hat in hand, humble, almost begging. (Transitive Oblique)

c. Officer McCoy, me and him was hassling and my gun went off, hitting him somewhere in his chest. (Intransitive)

d. thats the LAST time i use a non-digital camera when i'm doing serious photography because it saves all that ammoying hassle of SOD'S-BLOODY-LAW!!!!! (Nominal Mass)

e. I rarely paint my nails(It can be such a hassle!) (Nominal Count)

f. It's a very hassily event to do. I believe alot of reasons is it takes so much time, specially preperation. (Adjective Attributive)

g. She will not take part in Saturday's 5000m race, saying she is tired and bothered (Adjective Predicative)

h. However, we didn't have the time or the technical know-how to do this sort of hassling as the PDAs were ordered and the students were being briefed (Gerund)

Almost all the forms presented here subdivide into further formal variants, with different syntactic patterns for the verbal forms, grammatical number amongst the nouns, suffixation for the adjectives, as well as two gerund forms, one that maintains a verbal argument structure and another that adopts the nominal argument structure. However, these examples represent the overall pattern of formal variation. Table 1 summarises the relative number of occurrences of these grammatical classes and constructions.

Altogether some 16 different basic grammatical classes and constructions are found across the three lexemes in the dataset. The eight types given in Table 1 are the most important numerically and for the practical concern of data sparseness, the study is restricted to these forms.

Table 1. Principle Classes and Constructions of the Lexical Field BOTHER

Form	Dataset Occurrences
Count Noun *hassle*	146
Mass Noun *hassle*	217
Gerund *hassle*	40
Predicative Adjective *bother*	124
Intransitive *bother*	222
Transitive *annoy*	449
Transitive *hassle*	274
Transitive *bother*	275

The occurrences are annotated for a range of formal, semantic, and extralinguistic features. In total, some 120 features belonging to some 20 partially overlapping variables were analysed and tagged manually. At this level of onomasiological granularity and with only 2,000 occurrences, the formal variation in tense, aspect, mood, and post-predicate constituents did not reveal any informative variation in usage. There was some variation relative to person and number, but this was found to be an indirect result of other factors that we examine below. The nature of the corpus limits the range of extralinguistic variation that may be investigated. For this reason, the most insightful extralinguistic variable available for consideration is certainly the regional variation between American and British usage. This is stratified in the corpus and so straightforward to annotate. For the analysis of the synonymy *per se*, the semantic variables were the most informative and we will focus on these. Before we examine the variables in question, an important aside should be made.

Within corpus linguistics, there is a very reasonable tendency to avoid semantic feature analysis. This is for two reasons. Firstly, semantic annotation is largely manual. Such annotation entails a labour and time intensive process that limits considerably the number of observations that can be analysed and tagged. Since data sparseness is an ever-present problem in quantitative studies, this represents an inherent weakness that one wishes to avoid. Secondly, corpus linguistics, like all empirical methods, seeks to maximise objectivity. Semantic feature analysis is inherently subjective.

There are strong counterweights to these arguments. Although we can describe a great deal of linguistic structure limiting our research to formal phenomena, ultimately, especially within a framework such as Cognitive Linguistics, we must also apply these kinds of techniques to semantic struc-

ture. Although this will force us to work with smaller numbers of observations, it represents an inherent weakness of the method and it must be taken on board and considered when we estimate the value of the results it produces.

The same is true for the question of objectivity. We cannot pretend that any semantic analysis will be purely objective, but this should not stop us from investigating semantic structure. Quantitative studies of linguistic semantics simply repeat the kind of semantic analysis that traditional linguists use, but many hundreds of times. Although, in itself, this does not assure a higher degree of objectivity, the large number of examples does improve analytical reliability in a number of ways.

Firstly, by examining many hundreds, or thousands, of examples the researcher sees facets of usage that would not necessarily be found through hermeneutic reflection. Although this approach cannot hope to account for all possible uses, the analysis of large numbers of found examples offers the researcher an 'external', therefore objective, source for his or her analysis. However, this does not mean the analysis itself is more objective. Secondly, a quantitative and usage-based approach offers three means for result verification, which serve as check on the objectiveness of the analysis. In the first place, systematicity and intuitively sound patterns found by the statistical results are indications of accuracy in semantic analysis. It must be remembered that after the analysis, the results found through the statistical treatment of the data are independent of the researcher, and in this, are completely objective. When patterns of usage that match an intuitively sound perception of usage 'fall out' from the statistical analysis, we can be reasonably sure that the original semantic analysis is accurate. In a second place, confirmatory statistical techniques employ models of the data, based on the results of the analysis, to check their validity. If one may predict the usage of a word, in a given situation, to a very high level of accuracy, then we can be more sure that the original analysis is accurate. In a third place, one may repeat the analysis on a second dataset. If the results are comparable, then once again, we can be surer of the accuracy of the semantic analysis.

We concentrate on three semantic variables, the cause of the BOTHER event, the affect upon the patient of the event, and the presence or lack of humour in the description of the event. The annotation focuses not on the word, but on the entire utterance. In many cases, a great deal of context needs to be considered to accurately ascertain the cause or affect being described by the lexeme in question. Table 2 lists the three semantic variables and the features for which they are annotated.

Table 2. Semantic Features

Cause of Event	Affect on Patient	Humour
expenditure of energy imposition imposition / request request interruption condemnation tease aesthetics repetition	anger concern - thought emotional pain physical pain	Presence of humour Absence of humour

In order to avoid overlap between the variables, either the cause or the affect was coded, never both. Statistical techniques do not work when one has redundancy across variables. Certain cause features, for example, 'repetition', which systematically co-occurs with what would be the affect of 'boredom', are therefore a problem. Thus, for the purposes of the statistical analyses below, the cause and affect variables are treated as a single variable.

Most of the features should be self-explanatory, however several warrant a word of explanation. Three particularly important features include 'imposition', 'imposition-request', and 'request'. These features identify uses where the agent of the event imposes him or herself upon the patient or makes a request of him or her. Often, both these two features are present; when this is the case, the example is coded as 'imposition-request'. The clearest way to explain these features is by way of example. Examples (3a) - (3c) represent these semantic distinctions.

(3) a. While Valentine's Day is a nice thought, it's always such a hassle. Romance should never be an obligation, and neither should it be restricted to a single day, which are the messages Valentine's Day sends. (Imposition)

b. ... and walked up the Grays Inn Road being hassled by aggressive beggars who glared at me straight in the eyes, asking Got any change? (Imposition request)

c. I can then update the page, and won't need to hassle you for the results of matches that have been postponed. (Request)

The features 'aesthetics', 'condemnation', and 'tease' also deserve explanation. In the diary entries, speakers often experience BOTHER because

someone is judging them. This is quite distinct from a situation where classmates or friends are teasing the patient and also from a situation where some inherent quality in the world displeases the patient. Again, examples can clarify the semantic features in question as well as the kind of subtle semantic differences that the coding seeks to capture. A reasonably large amount of context was needed in order to accurately discern many of the semantic distinctions.

(4) a. "Now it's tough being an American. Everyone always gives us hassle for having a stupid president. Especially you Brits. You give us hassle for having a retard for a President. But we know he's a retard. (Condemnation)

b. bumping into Kath, which i always do when i'm fucked, and having lots of hugs. and not being able to pee in front of her in the toilets and hassling her because she has curly hair and i wanted to "ping" it. (Tease)

c. he dnt reilise tht she loves him sooo much it dnt bother her wot is on his face lol (Aesthetic)

It should also be stressed that 'humour' refers to the utterance in which the lexeme is used and to the intention of the speaker. The other features should be self-explanatory, their semantic distinctions being drawn in a similar manner to those described here.

3. Usage-Based Methodology. A Multifactorial Treatment of Results

3.1. Semantic Relations between Lemmata

Having completed the semantic analysis of the observations, we now have what are referred to as multiway contingency tables. These are three, four-way, or *n*-way tables of frequencies of co-occurring, extralinguistic, formal, and semantic features. Although one may not visualise a multiway table, the mathematical relations are simply the frequencies of co-occurrences of multiple features. These features are relative to various levels of granularity in the formal variation. For example, we can examine the correlation between the semantic variables and the three words without including the formal variation of each lemma. We can equally zoom in and examine the formal variation at a very fine-grained level, differentiating not only grammatical class and grammatical construction but also tense, mood, aspect,

and so forth. The limitation is data sparseness: as we include more detail in formal variation, the numbers of occurrences for each semantic feature drops quickly. At a certain point the frequencies of occurrences become too small for us to identify meaningful generalisations in the data.

Moreover, interpreting a three-way or four-way table of frequencies of co-occurrences is not possible without using multivariate tools. Exploratory techniques exist that search through these tables looking for patterns of correlations. In other words, mathematically, some features co-occur appreciably more often than others. In our case, these are the semantic features co-occurring with the various forms of *annoy*, *bother*, and *hassle*. One such exploratory technique is Correspondence Analysis. This simple statistical technique takes the frequencies of multiway tables and converts those frequencies to distances. It then conflates the multidimensional distances to a two-dimensional plane that maps the correlations between the features visually. Although this allows us to 'see' the correlations and differences between the forms and semantic features, one must be careful in reading such visualisations since, obviously, representing *n*-dimensions in a two-dimensional plane can be misleading. For this reason, the position of many of the data points relative to other data points can be misleading. Careful consultation and experience interpreting the plots is the only way to avoid misinterpretation.

Let us begin with a Bivariate Correspondence Analysis of the semantic variables relative to the three lemmata. Figure 1 is a correspondence map of the analysis. It should be remembered that relative proximity of the data points represents relative correlation.

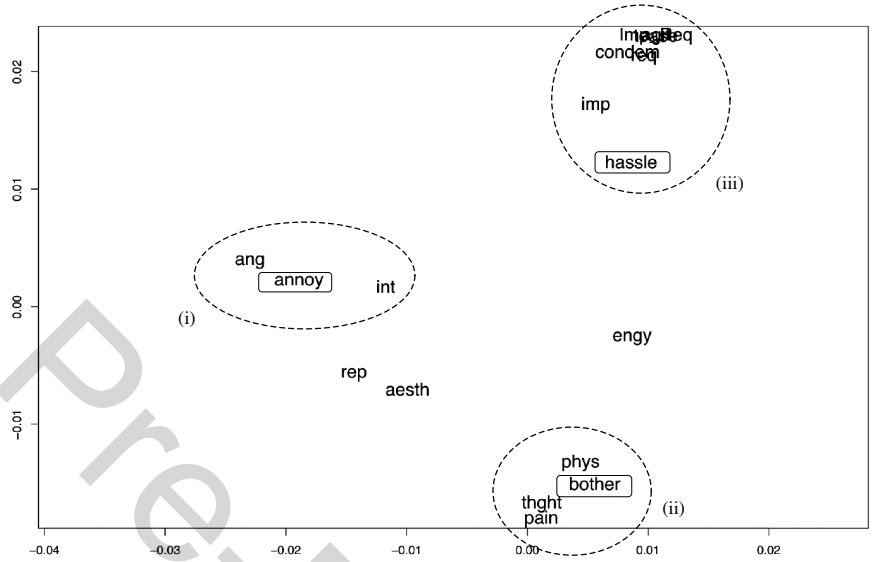


Figure 1. Correspondence Analysis BOTHER Lemmata and Cause-Affect

Interpreting the visualisations of Correspondence Analysis can be difficult. Let us move through a description of the plot, step by step. Firstly, on the left (i), we see *annoy*, grouped with 'anger' <ang> and 'interruption' <int>. The feature 'anger' <ang> is to the left of *annoy*, which stands between it and the other lemmata. The position of this feature shows that it is highly distinctive for the usage of the lemma *annoy*. This is intuitively sound: of the three lexemes in question, *annoy* represents the point of overlap with the concept of ANGER, an interpretation corroborated by traditional dictionaries. Also associated with the lemma *annoy* is 'interruption' <int>. However, the fact that this feature occurs to the right of the *annoy* data point, placed between the two other constructions, suggests that despite a clear association with *annoy*, this feature is shared to some extent by all three words.

Placed more or less evenly between (i) *annoy* and (ii) *bother*, we find two cause features, 'aesthetics' <aesth> and 'repetition' <rep>. We can suppose quite safely that these two features are characteristic of both these lemmata. The two features 'concern – thought' <thght> and 'emotional pain' <pain> lie just beneath the *bother* data point and so are distinctly associated with this lemma. Just as 'anger' is effectively unique to *annoy*,

the semantically similar features 'emotional pain' and 'concern – thought' are effectively unique to *bother*. This is also intuitively sound. A third feature, which was rare in the data, is also highly associated with the lemma *bother*. The cause feature 'physical pain' <phys> only occurs 10 times out of almost 2,000 observations. Of these 10 occurrences, 7 are with *bother*, 2 with *hassle*, and 1 with *annoy*. It seems with such small frequencies, we cannot draw any firm conclusions. However, in the dataset, to the extent that this feature occurs, it is associated with *bother*.

One of the three most important features in terms of frequency, occurring 650 times, is that of the 'expenditure of energy' <engy>. Its data point lies in the centre of the plot, equidistant from *hassle* and *bother*, yet relatively far from *annoy*. The position of this data point strongly suggests that this feature is characteristic of *bother* and *hassle*, more than of *annoy*.

Finally, the cluster in the top right (iii) sees *hassle* associated with a large number of overlapping semantic features. One feature, 'imposition' <imp>, is distinct from this micro-cluster and considerably closer to the data point of *hassle*. This may signify a stronger correlation but needs further verification. The dense cluster just above this point consists of request <req>, 'imposition request' <imp_req>, 'condemnation' <condemn>, and 'tease' <tease>. These four semantic features seem to identify two 'meanings' of the word, the 'imposition request' and simple 'request' features being semantically similar as well as the 'tease' and 'condemnation' features clearly carving out a similar semantic space.

We could not ask for clearer results in this first Correspondence Analysis . Each of the three lemmata are evenly dispersed across the plot, distinctly grouped by semantic features. Certain semantic features lie between the lemmata, showing overlap in the semasiological distribution. This kind of semantic map is a simple but powerful generalisation that shows the basic differences and similarities of usage across the three synonymous words.

At this point, it is worth noting that mapping the correlations between such semantic features and various forms should be seen as an indirect means for capturing the conceptual structure. The kind of the results we see here are intuitively sound and match the kind of results that one would posit using an individual's knowledge of a language. The important difference, of course, is that this technique permits repeat analysis, and is therefore easily verifiable.

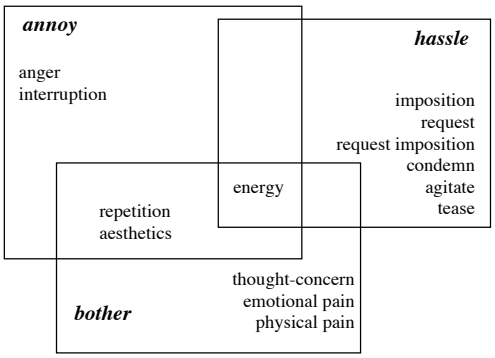


Figure 2. Box Summary BOTHER. Lemmata and Cause-Affect Features

We can summarise the results of the Correspondence Analysis with a box diagram. This is presented above in Figure 2. Although the box diagram adds nothing to the actual results, it is clear and more easily interpretable. Its downside is, by rendering the correlations discrete, it does not capture the semantic continua between the correlations.

Despite these intuitively attractive results, even dictionaries break down lemmata into grammatical classes and this kind of coarse-grain analysis is only helpful in mapping the aggregate meaning of the three words. Any accurate semantic description must look closer than this.

3.2. Grammatical Class, Grammatical Construction, and Semantic Similarity

Let us now repeat the analysis while rendering the formal dimension more fine-grained. Figure 3 plots a Correspondence Analysis that identifies correlation between cause-affect and class-construction.

In direct contrast to the lemma level of analysis, we see more semantic similarity between different words within the same class-construction than between the different forms of a single lemma. In group (i), we see how, relative to the semantic features in question, the transitive forms of *annoy* and *bother* group together. In contrast to this, the transitive use of *hassle* sees a distinct usage (ii), highly associated with instances of impositions and requests. Then a third group (iii) clusters the adjectival, nominal, and intransitive profilings of all three words.

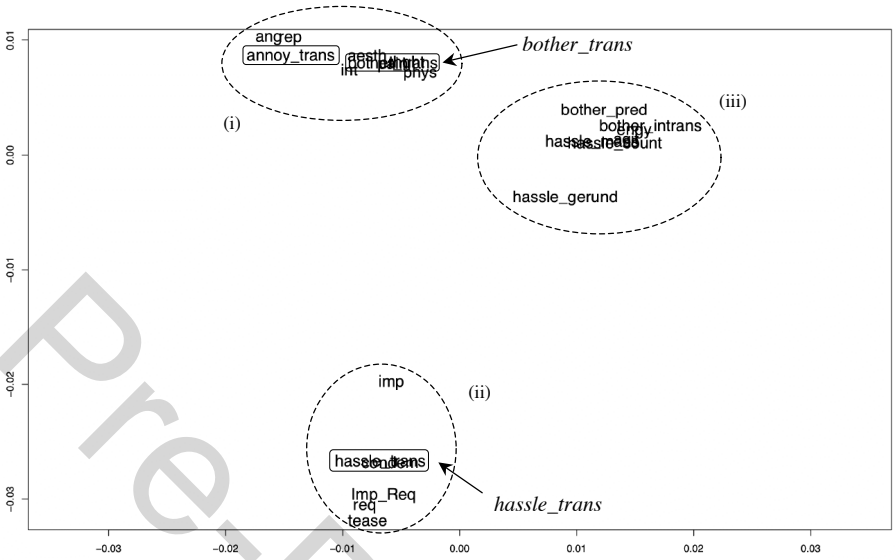


Figure 3. Correspondence Analysis of BOTHER. Class-Construction and Cause-Affect

Before we look more closely at the detail of these correlations, let us add another dimension to the analysis. Regional variation often has a profound effect on semantic variation. This is because even if a word or construction exists across all the varieties of a given language, this does not entail that it is used in the same manner. The countless 'false friends' between British and American English are testimony to this. However, if we distinguish the forms further, dividing between the British and American varieties, the analysis reveals an almost identical picture suggesting at this onomasiological level, there is little dialect variation. The plot in figure 4 visualises a Bivariate Correspondence Analysis of class-construction distinguished for dialect, correlated with the semantic features of cause and affect.

By splitting the class-constructions into British and American variants we double the number of forms, leading to a denser plot. Moreover, splitting the data offers two datasets for comparison. Assuming there is no substantial dialect variation, this serves as an indirect way of verifying the results. In light of this, the most important result of the Correspondence Analysis visualised in Figure 4 is that the three basic uses across the onomasiological field are maintained. Indeed in terms of placement and proximity, the map is little different to that given by the Correspondence Analy-

sis of the formal variation without the variable of dialect. The greatest difference in the results is that the outlying cause-affect features, with the exception of 'imposition' <imp>, have been 'brought into' the clusters. In the majority of cases, the dialectical pairs behave in the same manner. Only one pair splits between the different clusters; the Adjectival Construction for *bother*.

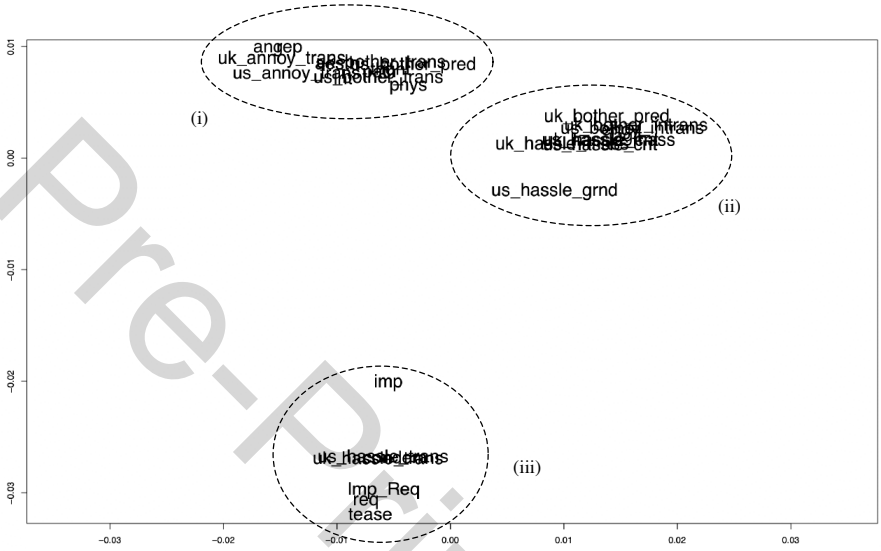


Figure 4. Correspondence Analysis of BOTHER. Class-Construction-Dialect and Cause-Affect

Let us look again, this time more closely, at the clusters. We can zoom in on each of the clusters identified in Figure 4 to see what features and forms are correlated.

Usage Cluster 1

Dialect Class Form

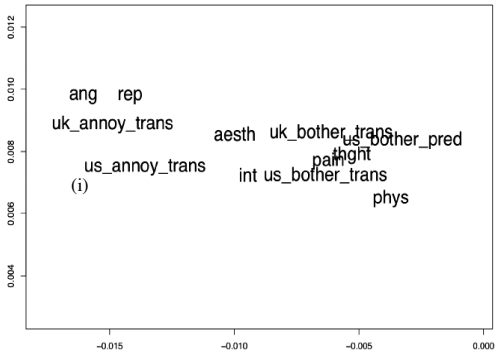
a. Transitive *annoy*
   Transitive *bother*

b. Am. Predicative *bother*

Affect Features

a. anger
   thought-concern
   emotional pain
   physical pain

b. repetition
   interruption
   aesthetics



The most surprising result here is that the American predicative form of *bother* has been clustered with these transitive forms. By dividing the words into two dialectally distinguished forms, we substantially reduce the number of co-occurrences with the various semantic features. This may mean that for a relatively infrequent form such as the predicative *bother*, the results are erroneous. We will assume the accuracy of the correspondence analysis, but in this case, further investigation is necessary.

The two transitive forms of *bother* and *annoy* cluster with what seem to be two sets of similar semantic features. Firstly, there appears to be a semantic cline from the affect of 'anger' through 'emotional pain' and 'thought-concern' to perhaps 'physical pain'. The similarity of these semantic features suggests a clear 'meaning' is associated with these two forms. Moreover, the systematicity represented by the grouping of these semantic features adds weight to the argument that the analysis and annotation has successfully operationalised the subjective nature of these features.

The second sub-group of semantic features found here is less homogenous, but still reasonably coherent. This group, in contrast to the other features, includes causes that are of a relatively inconsequent nature. Causes such as 'repetition', 'interruption', or 'aesthetic displeasure' are similar in that they are little more than inconveniences for the patient.

The kind of usage in question can be explained by way of example. The 'anger', 'thought-concern', and 'emotional pain' uses of the transitive *annoy* and transitive *bother* are represented by examples (5a) – (5c). This is contrasted by examples (6a) – (6b), which are typical of causes such as 'interruption' and 'aesthetic displeasure'.

(5) a. There are even people out there that annoy the hell out of me. (Anger)

b. they can get 2 fuk.. im not gona let it bother me.. (Thought-concern)

c. It bothers me when I am starting to beg for people to think about me when I've never done it before. I cannot explain how I feel right now. (Emotional pain)

(6) a. oh on the last night the guys kept annoying him while he was trying to sleep (Interuption)

b. Ok, I don't really like my mood theme. I love Nightmare and all but the theme is bothering me for some reason. (Aesthetic)

Usage Cluster 2

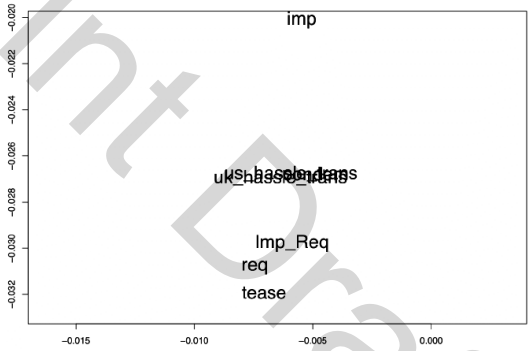
Dialect Class Form

Brit. Transitive *hassle*
Am. Transitive *hassle*

Affect Features

a. condemnation
tease

b. imposition
request
imposition-request



Here, we see that the transitive form of *hassle* stands out as a relatively unique usage. It is associated with two very clearly grouped sets of semantic features. Again the systematicity of the semantic feature groupings strongly supports the success of this variable's analysis and annotation. These groups include, on the one hand, 'tease' – 'condemnation' and on the other hand, 'imposition' – 'request' – 'imposition-request'. It seems that

this form is distinct in its usage and possesses two relatively distinct meanings. Examples (7) and (8) represent the two clusters of features and the kind of usage they indicate.

(7) a. Some smokers also have a dream that someday the nonsmoking world will quit hassling them about their smoking. (Condemnation)

b. Anyway today nothing excited happen excpet all my teachers had to be taught to be better teachers (their turn to be bored) and I had substitutes in every class. I hassled them and had fun doing it! (Tease)

(8) a. i saw him yesterday and he was being all touchy feely.....i don't want him back...but hes hassling me now and I fee sorry for liz(his new g/f) (Imposition)

b. she had other ideas and hassled Dave to walk her to the train station (Request)

c. Ford and Greg: Nah, the real Glasgow neds hassle us for our wallets (Imposition-request)

Usage Cluster 3

Dialect Class Form

a. Intransitive *bother*

b. Mass *hassle*
Count *hassle*
Gerund *hassle*

c. Brit. Predicative *bother*

Affect Features
energy
agitation

The third usage cluster of correlations includes the nominal-gerundive forms as well as the intransitive forms. Before the addition of the variable of dialect, it also included the adjectival forms. First, it must be noted that a wide range of forms are grouped relative to only two semantic features, 'expenditure of energy' <engy> and 'agitation' <agit>. Second, the first of these two semantic features is the most common of the dataset and the sec-

ond is a relatively infrequent feature. From this, we can tentatively deduce that in fact the non-verbal forms are associated with the 'expenditure of energy' relative to the verbal forms, which represent a semantically more complicated profiling. The correlation with the feature of 'agitation' is likely to be incidental.

Lastly, the British form of the predicative remains in this cluster where it was before we added the variable of dialect. This is in contrast to the American predicative form, which as we saw, is found in Cluster 1. However, by adding the variable of dialect, we increase the number of correspondences calculated by the analysis considerably. For a relatively infrequent form, such as the predicative *bother*, we are faced with a degree of data sparseness. It is therefore possible that the results presented in Figure 3, are misleading. If this were the case, it would leave all the non-verbal uses together and associate them with the single most common semantic feature, the 'expenditure of energy'. Further investigation is needed in order to determine if there is a distinction in use between the dialects and whether this adjectival form does, in fact, divide along the lines suggested by the analysis.

Let us add one last variable, that of 'humour'. For such negative emotion terms as *annoy*, *bother*, and *hassle*, this feature is clearly marked. It is important since it captures a difference that further distinguishes one of the forms, transitive *hassle*. In Figure 5, the most striking feature is that the clustering captured by the analysis remains stable after the addition of the extra variable. This further re-assures us that the analysis is capturing real semantic structures extant in language use. However, the feature itself proves to be important. The lack of 'humour' <NHum> falls squarely between both the transitive *bother* - *annoy* cluster and the nominal-adjectival-intransitive cluster contrasted starkly by the clear correlation between the presence of 'humour' <Hum> and the transitive *hassle* uses. Example (9) captures the kind of uses in question.

(9) a. Vicky spent most of the days hassling cows and sheep. Occasionally she would do a little skip or run for no reason

b. ... sitting outside Mcdonalds and hassling kids for change, and taxing people. The west end is the Crewe chav centre, other wise known as "The Cronx".

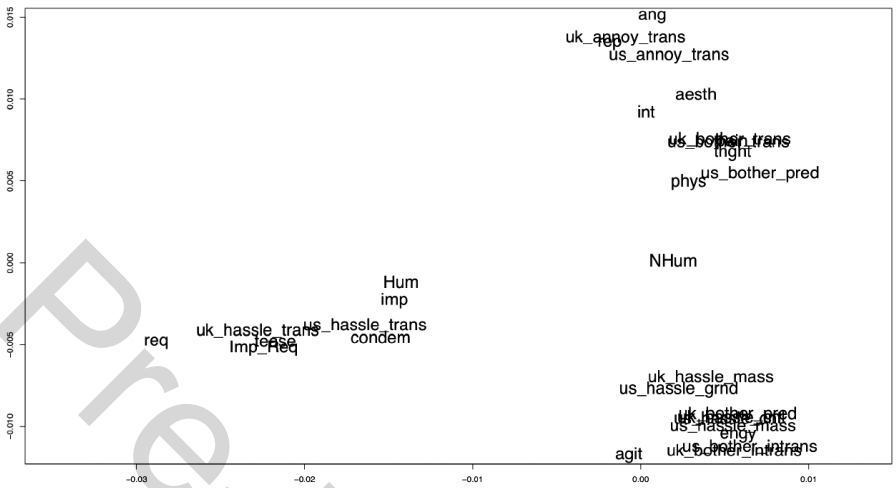


Figure 5. Multiple Correspondence Analysis of BOTHER. Class-Construction-Dialect, Cause-Affect, and Humou*r*

We can perform one last statistical analysis to help verify our findings. Hierarchical Cluster Analysis functions in a similar way to Correspondence Analysis, converting frequencies to distances. However, instead of plotting those distances, it uses a pre-determined distance measure to identify clusters. The visualisation takes the form of a dendogram. This does not show what semantic features cause the clustering of the forms, but it does offer a clearer picture and allows us to include significance testing via bootstrap resampling. Bootstrapping is a complicated mathematical procedure for determining the probability that a given result will be repeated, given the same data. In the plot below, the different forms are clustered relative to the semantic features cause-affect and humour.

The results clearly verify the results of the Correspondence Analysis. Not only are the same clusters identified, a further more subtle distinction is added. Although the intransitive forms, adjectival, and nominal-gerund forms are grouped together, they are once again subdivided.

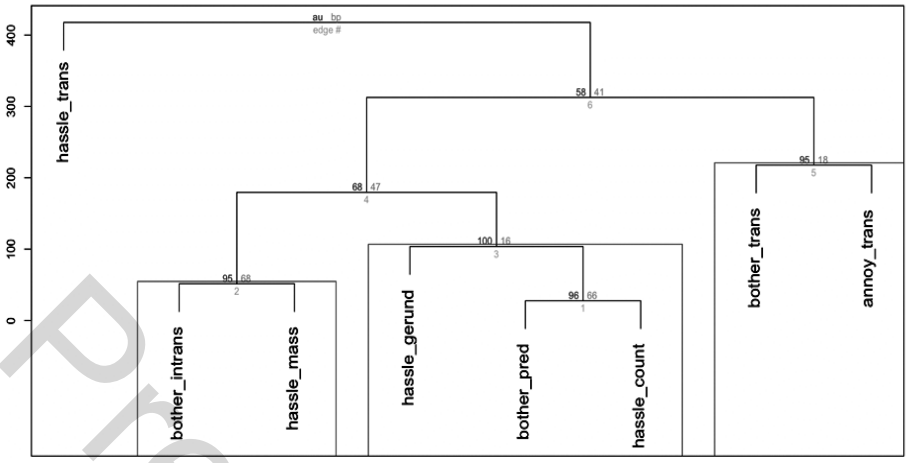


Figure 6. Hierarchical Cluster Analysis (Ward) BOTHER. Class-Construction Cause-Affect

In the plot, the boxes drawn around the dendogram clusters are the bootstrapping results. Two different bootstrapping algorithms are used. The numbers at the top of the boxes represent the results of the bootstrap samples, the first number is the results of the more reliable multiscale bootstrap sample and the second number the simpler and less reliable normal bootstrap. The closer the figure is to 100, the better the result. In terms of probability we have excellent results that strongly suggest these clusters are accurate representations of the data.

Note that the Cluster Analysis identifies a distinction that is not apparent in the Correspondences Analyses. What was referred to as cluster 3 above, is here subdivided into two sub-clusters: intransitive *bother* and mass noun *hassle* on the one hand versus gerund *hassle*, count noun *hassle,* and adjectival *bother* on the other. Investigation into this distinction is beyond the scope of the current study, but the Cluster Analysis suggests that there is a clear usage difference between these two groups. Most importantly, the bootstrapping on the Cluster Analysis offers us a means of verification for the results found in the Correspondence Analysis. It shows that there is an extremely high probability that if we repeated this study many hundreds of times, we would obtain the same groupings of form and usage.

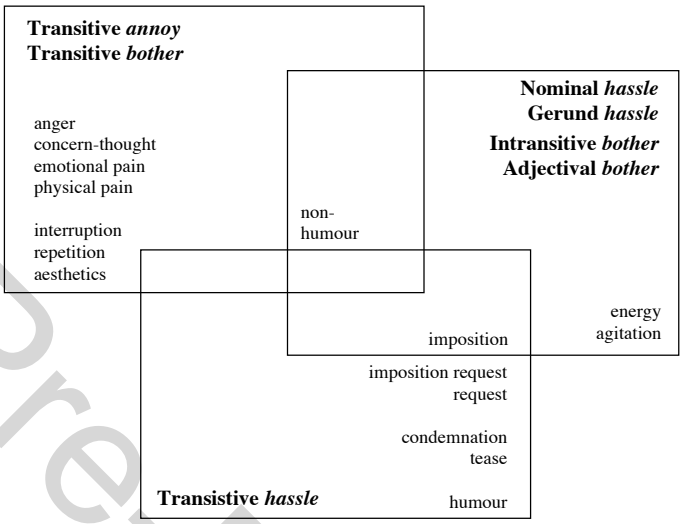


Figure 7. Box Summary BOTHER. Class-Construction and Cause-Affect and Humour

By way of conclusion, Figure 7 presents a box summary of the findings. The results, when summarised in this manner, resemble the conceptual space maps of the Structuralist era. However, the results presented here fall out from a mathematical logarithm that examines frequencies of co-occurring features of language use. This does not at all prove the results, indeed far from it, nor does it necessarily mean they are more accurate. However, it does mean that the analysis is repeatable. This can be done with similar data from the same corpus to verify that this is indeed an accurate depiction of the semantic structure associated with the three words for this kind of language. However, this verification can also be performed with different corpora of different kinds of language to determine to what degree the results are influenced by the register and mode of the language rather than the lexical semantic structure *per se*. These possibilities for verification are an important addition to Cognitive Semantic analysis, especially since this method can be expanded to more culturally rich concepts.

Despite the fact that the discrete boxes used to summarise the results of the Correspondence Analyses may be misleading in their simplicity, they do help appreciate how, via the careful semantic annotation of some 2,000

examples, quantitative investigation helps us map semantic structure. The diagram can be seen as a summary of the conceptual associations of different yet similar linguistic forms. By adding other semantic features, such as 'agent type' and 'patient type', 'topic of discourse', as well as more formal detail, such as variations in post-predicate argument structure and so forth, we could enrich this map, adding finer levels of granularity of formal and semantic detail. For this, perhaps extra examples would be needed since the more factors one considers simultaneously, the more data one requires. Nevertheless, this small study has hopefully shown how quantitative techniques can capture semantic similarity between words and do so while accounting for some of the multidimensionality of language.

4. Summary

This study has successfully made four points. Firstly, we have seen how quantitative and multidimensional techniques can help map usage patterns, patterns that theoretically represent the grammar of that language. In this way, we have seen how we can vary the level of granularity of the study by increasing the degree of formal details considered, contrasting a study at the level of the lemma with a study at the level of grammatical class and construction. Secondly, we have seen how it is possible to use direct semantic analysis in quantitative approaches. The semantic features in question may be determined subjectively, but the systematicity and intuitively coherent results demonstrate that careful analysis and annotation of even subjective semantic characteristics of language use is operationalisable. Thirdly, we saw how a simple statistical technique, Correspondence Analysis, can help capture the multidimensional correlations produced by the semantic analysis. Although the discussion did not directly compare Correspondence Analysis with other techniques that have been used to describe synonymous relations, the technique proved successful. Fourthly and returning to the first point, we have seen how the study of synonymy and semantic relations of similarity can be used to posit hypothetical conceptual structures. Since we argue that usage is conceptually motivated, the patterns in usage represent more than grammar, but the conceptual structures argued by Cognitive Linguistics to motivate grammar. Quantitative usage-based studies of this kind, therefore, offer an indirect yet verifiable approach to the study of conceptual structure.

There are, of course, certain deductions that this study cannot draw. Firstly, we are in no place to make hypotheses about the categorisation of the concepts. It may well be that in these instances, the frequency data do represent prototype effects and category structure, but until we understand the relationship between ontological salience and frequency, this is an assumption we cannot make. Secondly and similarly, we cannot draw any conclusions about the cognitive salience and the processing of the lexical semantics and its integration with the grammatical semantics. At this level, corpus-driven research must pass the torch to psychological experimentation, for its frequency counts offer few insights.

To the extent that the corpus is representative of language and to the extent that the dataset is representative of the corpus, we can propose a partial semantic map of the lexical encoding of the concept BOTHER. There are other words and expressions that should be included, just as different registers and modes of language, and so we cannot say that we have fully described the synonymy of these words or the conceptual structure they are used to represent. However, we have a partial map of the patterns of language use, patterns we argue indicate conceptual structure.

The next step will be to test these findings. This needs to be done at two levels. Firstly, new data from a different sample of language need to be analysed and the results compared. Secondly, confirmatory statistical techniques need to be used to demonstrate that for the datasets in question, the results are more than chance and do map, or model, the reality of the data. Perhaps in comparison to other methods of language analysis, these results seem conditional and limited. Even if this is true, the results are verifiable and are truly usage-based representations of the linguistic patterns that make up the grammar of a language.

Notes

1. Note that both authors have since stepped back from the stronger claims made in this vein. For more recent discussion on the relationship between frequency based evidence and cognition, see Glynn (2006, in press), Schmid (2007), and Gilquin (2008).
2. All examples are taken from a corpus built from on-line personal diaries. The details of which are given in section 2.2.
3. Further discussion concerning these lines of research and the methods used may be found in Tummers et al. (2005) and Heylen et al. (2008).

4. Glynn (2004, 2009) goes further to argue that lexical study is not at all possible without morpho-syntactic context. It is argued that grammatical semantics are inherently interwoven with lexical semantics and, regardless of redundancy, the only way to explain lexical structure is by simultaneously accounting for grammatical structure.
5. The importance of extralinguistic factors in Cognitive Linguistics is gaining wide acceptance. See Geeraerts (1995), Kristiansen & Dirven (2008), Geeraerts et al. (forthc.) for discussion and examples of this line of research.

References

Croft, William
2001 *Radical Construction Grammar, Syntactic Theory in Typological Perspective*. Oxford: OUP.

Croft, William
2008 Toward a social cognitive linguistics. V. Evans & S. Pourcel (eds), *New Directions in Cognitive Linguistics*, 395-420. Amsterdam: Benjamins.

Dąbrowska, Ewa
2006 Low-level schemas or general rules? The role of diminutives in the acquisition of Polish case inflections. *Language Sciences* 28: 120-135.

Dirven, René. Goossens, Louis. Putseys, Yves. & Vorlat, Emma
1982 *The scene of linguistic action and its perspectivization by SPEAK, TALK, SAY, and TELL*. Amsterdam: Benjamins.

Divjak, Dagmar
2006 Delineating and Structuring Near-Synonyms. St. Th. Gries & A. Stefanowitsch (ed.), *Corpora in cognitive linguistics*, 19-56. Berlin: Mouton.

Divjak, Dagmar & Gries, Stefan. Th.
2006 Ways of trying in Russian: clustering behavioral profiles. *Journal of Corpus Linguistics and Linguistic Theory* 2: 23-60.

Fillmore, Charles
1977 Topics in Lexical Semantics. R. Cole (ed.), *Current Issues in Linguistic Theory*. 76-138. Bloomington: Indiana University Press.

Fillmore, Charles
2003 *Form and Meaning in Language*. Standford: CSLI.

Geeraerts, Dirk
2005 Lectal data and empirical variation in Cognitive Linguistics. F. J. Ruiz de Mendoza Ibañez & S. Peña Cervel (eds), *Cognitive*

Linguistics. Internal Dynamics and Interdisci plinary Interactions, 163-189. Berlin: Mouton.

Geeraerts, Dirk
2006 Methodology in Cognitive Linguistics. G. Kristiansen, M. Archard, R. Dirven, & F. J. Ruiz (eds), *Cognitive Linguistics: Current Applications and Future Perspectives*. 21-49. Berlin: Mouton.

Geeraerts, Dirk Grondelaers, Stefan. & Bakema, Peter
1994 *The Structure of Lexical Variation. Meaning, Naming, and Context*. Berlin: Mouton.

Geeraerts, Dirk, Kristiansen, Gitte, & Peirsman, Yves (eds)
Forthc. *Advances in Cognitive Sociolinguistics*. Berlin: Mouton

Gibbs, René
2007 Why cognitive linguists should care more about empirical methods. M. Gonzalez-Marquez I. Mittelberg, S. Coulson, & M. Spivey (eds), *Methods in Cognitive Linguistics*, 2-18. Amsterdam: Benjamins.

Gilquin, Gaëtanelle
2008 What you think ain't what you get: highly polysemous verbs in mind and language. J.-R. Lapaire, G. Desagulier, & J.-B. Guignard (eds), *From Gram to Mind. Grammar as Cognition*, 237-258. Bordeaux: Presses universitaires de Bordeaux.

Glynn, Dylan
2002 Love and Anger. The grammatical structure of conceptual metaphors. *Style. Cognitive Approaches to Metaphor* 36:541-559.

Glynn, Dylan
2004 Constructions at the Crossroads. The Place of construction grammar between field and frame. *Annual Review of Cognitive Linguistics* 2:197-233.

Glynn, Dylan
2005 Concept Delimitation and Pragmatic Implicature. Issues for the study of metonymy. K. Kosecki (ed.), *Perspectives on Metonymy*, 157-174. Frankfurt: Lang.

Glynn, Dylan
2006 Iconicity and the Grammar - Lexis Interface. E. Tabakowska, C. Ljungberg & O. Fischer (eds), *Iconicity in Language and Literature 5*, 267-286. Amsterdam: Benjamins.

Glynn, Dylan
2007 Rain and Snow in West Germanic. A test case for cognitive grammar. J.-R. Lapaire, G. Desagulier, & J.-B. Guignard (eds), *From Gram to Mind: Grammar as Cognition*, 191-212. Pessac: Presses Universitaires de Bordeaux.

Glynn, Dylan.
2008 Arbitrary Structure, Cognitive Grammar, and the partes orationis. A Study in Polish Paradigms. K. Willems & L. De Cuypere (eds),

Naturalness and Iconicity in Linguistics, 215-239. Amsterdam: Benjamins.

Glynn, Dylan
2009. Polysemy, Syntax, and Variation. A usage-based method for Cognitive Semantics. V. Evans & S. Pourcel (eds). *New Directions in Cognitive Linguistics*, 77-105. Amsterdam: Benjamins.

Glynn, Dylan
Forthc. The semantics of extralinguistic variation. A quantitative study of dialect effects on polysemy. D. Geeraerts, G. Kristiansen, & Y. Peirsman (eds), *Advances in Cognitive Sociolinguistics*. Berlin: Mouton.

Gries, Stefan. Th.
1999 Particle movement: a cognitive and functional approach. *Cognitive Linguistics* 10:105-45.

Grondelaers, Stefan. Speelman, Dirk & Geeraerts, Dirk
2007 A Case for a Cognitive Corpus Linguistics. M. Gonzalez-Marquez I. Mittelberg, S. Coulson, & M. Spivey (eds), *Methods in Cognitive Linguistics*, 149-170. Benjamins: Amsterdam.

Heylen, Kris
2005 A Quantitative Corpus Study of German Word Order Variation. Stephan Kepser & Marga Reis (eds.), *Linguistic Evidence: Empirical, Theoretical and Computational Perspective*, 241-264. Berlin: Mouton.

Heylen, Kris Tummers, José, & Geeraerts, Dirk
2008. Methodological issues in corpus-based Cognitive Linguistics. G. Kristiansen & R. Dirven (eds), *Cognitive Sociolinguistics: Language Variation, Cultural Models, Social Systems*, 91-129. Berlin: Mouton.

Kristiansen, Gitte & Dirven, René (eds)
2008 *Cognitive Sociolinguistics: Language Variation, Cultural Models, Social Systems.* Berlin: Mouton.

Langacker, Ronald
1987 *Foundations of Cognitive Grammar*. Vol. 1, *Theoretical Prerequisites*. Stanford: Stanford University Press.

Lehrer, Adrienne
1982 *Wine and Conversation*. Bloomington: Indiana University Press.

Newman, John & Rice, Sally
2004a Patterns of usage for English SIT, STAND, and LIE: A cognitively-inspired exploration in corpus linguistics. *Cognitive Linguistics* 15: 351-396.

Newman, John & Rice, Sally
2004b Aspect in the making: A corpus analysis of English aspect-marking prepositions. S. Kemmer & M. Achard (ed.), *Language, Culture and Mind*, 313-327. Stanford: CSLI.

Rudzka-Ostyn, Brygida
1995 Metaphor, Schema, Invariance: The Case of Verbs of Answering. L. Goossens**,** P. Pauwels, B. Rudzka-Ostyn, A.-M. Simon-Vandenbergen & J. Vanparys (eds), *By Word of Mouth. Metaphor, metonymy and linguistic action in a cognitive perspective*, 205-244. Amsterdam: Benjamins.

Schmid, Hans-Jörg
1993 *Cottage, idea, start: Die Kategorisierung als Grundprinzip einer differenzierten Bedeutungsbeschreibung*. Tübingen: Niemeyer.

Schmid, Hans-Jörg
2000 *English Abstract Nouns as Conceptual Shells. From Corpus to Cognition*. Berlin: Mouton.

Schmid, Hans-Jörg
2007 Entrenchment, Salience, and Basic Levels. D. Geeraerts & H. Cuyckens (eds), *The Oxford Handbook of Cognitive Linguistic*, 117-138. Oxford: OUP.

Speelman, Dirk & Geeraerts, Dirk
Forthc. Causes for causatives. The case of Dutch *doen* and *laten*. *Linguistics of Causality.* T. Sanders & E. Sweetser (eds). Cambidge: Cambridge University Press

Talmy, Leonard
1988 The relation of grammar to cognition. B. Rudzka-Ostyn (ed.), Topics in cognitive linguistics. Amsterdam: Benajmins.

Tummers, José, Heylen, Kris & Geeraerts, Dirk
2005 Usage-based approaches in Cognitive Linguistics: A technical state of the art. *Corpus Linguistics and Linguistic Theory* 1: 225-26.