Corpora and Cognitive Linguistics

Corpora e linguística cognitiva

John Newman*
University of Alberta
Edmonton / Canada

ABSTRACT: Corpora are a natural source of data for cognitive linguists, since corpora, more than any other source of data, reflect "usage" – a notion which is often claimed to be of critical importance to the field of cognitive linguistics. Corpora are relevant to all the main topics of interest in cognitive linguistics: metaphor, polysemy, synonymy, prototypes, and constructional analysis. I consider each of these topics in turn and offer suggestions about which methods of analysis can be profitably used with available corpora to explore these topics further. In addition, I consider how the design and content of currently used corpora need to be rethought if corpora are to provide all the types of usage data that cognitive linguists require.

KEYWORDS: corpora, cognitive linguistics, metaphor, polysemy, synonymy, prototypes, constructional analysis, statistics, R statistical program

RESUMO: Corpora são uma fonte natural de dados para a linguística cognitiva, uma vez que, estes, mais que qualquer outra fonte de dados, refletem "o uso" – a noção que é frequentemente apontada como tendo importância crítica para o campo da linguística cognitiva. Corpora são relevantes para todos os principais tópicos de interesse da linguística cognitiva: metáfora, polissemia, sinonímia, protótipos e análise construcional. Neste artigo, considerarei cada um desses tópicos e oferecerei sugestões sobre quais os métodos de análise podem ser utilizados com os corpora disponíveis para melhor se explorarem esses tópicos. Adicionalmente, discuto como a arquitetura e o conteúdo dos corpora atualmente disponíveis necessitam ser repensados se pretenderem oferecer todos os tipos de dados de uso necessários às análises da linguística cognitiva.

PALAVRAS-CHAVE: Corpora, linguística cognitiva, metáfora, polissemia, sinonímia, protótipos, análise construcional, estatística, programa estatístico R

* john.newman@ualberta.ca

1. Cognitive Linguistics

In its broadest sense, *cognitive linguistics* is concerned with general principles that provide some explanation for all aspects of language, including principles drawn from disciplines other than linguistics (cf. EVANS; GREEN, 2006, p. 27-28). Any intellectual movement which attempts to be so all-encompassing in its scope and so multi-disciplinary in its approach must inevitably lead to a proliferation of data types, methodologies, and strategies of persuasion. One may not be convinced that the field of cognitive linguistics will achieve all the goals it has set itself–which intellectual movement has?–but one can be grateful to the rise of cognitive linguistics for the balance it brings to the study of language, offering linguists a more rounded and more complete agenda for research than the relatively circumscribed, self-absorbed, self-referential, inward-looking kind of theorizing which constituted mainstream linguistic research, at least in syntax, in the latter half of the twentieth century.

In a reflective critique of the current state of cognitive linguistics, especially its methodologies, Geeraerts (2006, p. 29) refers to the "growing tendency of Cognitive Linguistics to stress its essential nature as a usage-based linguistics", a tendency which points unequivocally in the direction of corpus-based research. Certainly, there are now important collections of papers exemplifying corpus-based methods for a cognitive-linguistic audience, e.g., Gries and Stefanowitsch (2006), Stefanowitsch and Gries (2006), and Lewandowska-Tomaszczyk and Dziwirek (2009). Geeraerts choice of words–"growing tendency"–hints, however, at the unsettled role of usage-based methods in the field. On the one hand, it is sometimes claimed that such methods are a crucial component of a cognitive linguistic approach (cf. EVANS; GREEN, 2006, p. 108). Clearly, however, not everyone who purports to be a cognitive linguist has seen usage-based methods in quite the same way. If they did, there would not be a "growing tendency" to rely on such; instead, the use of these methods would be firmly entrenched in the practice of cognitive linguistics.

In the following sections, I comment on a variety of ways in which corpora can be exploited in the study of topics which have been central in the field of cognitive linguistics: metaphor, synonymy, polysemy, prototypes, and constructional analysis. Throughout, the emphasis will be on methods which I consider most promising in terms of their potential to yield interesting

results.[1] The Appendix contains the R scripts (see The R Project for Statistical Computing) I relied on to calculate the statistical measures and visualizations of the data. I will also comment briefly on the kinds of corpora which cognitive linguists should give more attention to.

2. Metaphor

A major advance in research on metaphor, and a prominent part of the cognitive linguistic movement, has been the adoption of a conceptually based approach to understanding metaphor. If ARGUMENTS ARE STRUCTURES and LOVE IS WAR, it is the concepts of argumentation, war, love etc. which are at the heart of these metaphors, not the words *argument*, *war*, *love* etc. This reconceptualization of metaphor research opened up fascinating new avenues of research. While acknowledging the conceptual breakthrough behind this development, I suggest that it is now appropriate to broaden the scope of inquiry into metaphor by appealing to more usage-based methods, in particular corpus-based methods, than has been customary. Taking metaphor research in this direction is, in fact, perfectly in keeping with the increasing empiricization of the field. My views on this seem to be simply echoing views already expressed by some researchers within the field. Ray Gibbs, for example, explicitly endorses a greater role for corpus-based research on metaphor in the following remarks:

> But I am most impressed these days with the development of work on corpus linguistics and conceptual metaphor. The research using corpora is important because it forces scholars to be more explicit on the procedures used for identifying metaphor in both language and thought, which is a necessary complement to more traditional introspectionist cognitive linguistic work on conceptual metaphor. (GIBBS, quoted in VALENZUELA, 2009, p. 310)

Later in the same interview, Gibbs calls for more attention to lexical and grammatical behaviors associated with metaphorical usage, which I, too, would advocate.

What might a corpus-based approach to metaphor entail? I see the insights of the cognitive linguistic research as providing a natural starting point for a corpus-based approach, in so far as these insights have identified a host of

[1] I am grateful to Ewa Dąbrowska for sharing the data on English motion verbs in Table 3 and Dagmara Dowbor for sharing her complete data on *over*. I am also grateful for comments by anonymous reviewers on an earlier draft of this article.

metaphorical mappings from source to target domains which in turn should stimulate research into the usage of these metaphors. ARGUMENTS ARE STRUCTURES may be a possible metaphorical mapping, but, compared with other metaphorical mappings, how often does this particular mapping actually occur? Or, how often do argument events show such structure (always, nearly always, hardly ever, etc.)? When discourse participants rely on that metaphor, which lexical items and constructions, as a matter of fact, constitute the vehicle of expression of the metaphor? Which conceptual mappings are more entrenched? We have become accustomed to rethinking syntactic concepts like ditransitive structures, passive structures etc. in terms of actual usage of such structures. In a similar way, it is appropriate to rethink the ARGUMENTS ARE STRUCTURES conceptual mapping, and others like it, in terms of usage. Usage of some metaphors more than others may lead us to insights into processing biases some of which may reflect fundamental aspects of human conceptualization while others may reflect culturally specific preferences. Investigating the degree of entrenchment of a metaphorical structure, established through the study of corpus-based usage, seems the natural next step to take if we are interested in such questions.

The presence of metaphorical usage in a corpus is clearly not something that is easily ascertained, unless the corpus has already been annotated to facilitate such searches.[2] An example of such a corpus would be one which has been annotated for word senses along the lines of WordNet (FELLBAUM, 1998) and related projects such as Euro WordNet (VOSSEN, 1998) and Global WordNet.[3] In WordNet, various senses of a word receive different *sense keys* which identify the sub-senses of a word and a set of semantically similar words, a *synset*, may share the same sense key. So, for example, *war* and *warfare* constitute one synset, sharing a sense key of 1:04:02 representing the shared sense of an active struggle between competing entities. In this sense key, "1" represents the syntactic category of noun, "04" denotes nouns representing acts or actions, and "02" is a unique identifier of the (shared) sense of the lemma

[2] See Philip (in press) for a helpful and critical review of various semi-automated approaches to identifying metaphors in a corpus. In Philip's own approach (2010, in press) *keywords* are first identified (where *keywords* are understood as words which are more common in a statistically significant sense in one corpus than in a reference corpus). Further procedures are followed, focusing on the *low-frequency* content words among the key words. See also Fass (1991) for an interesting computer-based approach to discriminating between metonymy and metaphor.

[3] See <http://www.globalwordnet.org/>.

war and the lemma *warfare*. In Table 1 below, we see three different sense keys associated with the pair *war* and *warfare*, the senses they represent, and an example of use of each.

TABLE 1
Selected WordNet sense keys and uses of *war* and *warfare*

	WORD	SENSE KEY	SENSE	EXAMPLE
1	*war*	1:04:00	the waging of armed conflict against an enemy	*thousands of people were killed in the* ***war***
2	*war*	1:04:01	a concerted campaign to end something that is injurious	*the* ***war*** *on poverty*
3	*war*	1:04:02	an active struggle between competing entities	*a* ***war*** *of wits*
4	*warfare*	1:04:02	an active struggle between competing entities	*Diplomatic* ***warfare***

It is clear how useful this kind of annotation would be if searching for metaphorical uses of both *war* and *warfare* – one would search these words when they occur with sense keys 1:04:01 or 1:04:02. It should be noted, though, that simply searching on these number sequences alone will not uniquely identify metaphorical uses. Other synsets of nouns relating to actions can utilize these same numeric sequences for other kinds of senses. For example, *battle* has senses utilizing these sequences where they indicate the senses shown in Table 2. In this case, the 1:04:01 sense is arguably the metaphorical use and 1:04:02 the more literal use. The sense key establishes a unique sense in relation to lemmas within a synset, but not between synsets. Thus, even with a WordNet annotated corpus, there is no simple search that will identify all metaphorical meanings relating to WAR.[4]

[4] Examples of corpora annotated for English WordNet are SemCor 3.0, based on the BROWN corpus (<http://www.cs.unt.edu/~rada/downloads.html#omwe>) and the Princeton WordNet Gloss Corpus of 1.6 million words (<http://wordnet.princeton.edu/glosstag.shtml>). There is substantial work involved in annotating a corpus for WordNet senses (as there is for most kinds of semantic annotation) which helps explain the lack of wide-spread availability of such corpora. A promising approach to a practical solution for adding WordNet annotations is found in Stamou, Andrikopoulos, and Christodoulakis (2003) where the authors describe a module WnetTag which retrieves and displays all

TABLE 2
Selected WordNet sense keys for *battle*

	WORD	SENSE KEY	SENSE	EXAMPLE
1	*battle*	1:04:01	an energetic attempt to achieve something	*he fought a **battle** for recognition*
2	*battle*	1:04:02	an open clash between two opposing groups (or individuals)	*police tried to control the **battle** between the pro- and anti-abortion mobs*

Another way of assigning semantic categories to words, though less complete than WordNet in the range of semantic relations to be indicated, is the "UCREL semantic analysis system" (USAS).[5] USAS relies on a classification into broad categories represented by the letters of the alphabet and narrower sub-categories indicated by additional delimiting numbers. For example, the category of government and the public domain is the G category, G2 is crime, law and order, G2.2 is general ethics. There is also a category of names and grammatical words that is assigned to words traditionally considered to be empty of content (i.e., closed class words) and proper nouns. USAS has been developed with automatic semantic tagging in mind and a detailed description of the tagging process and the array of sub-routines required to effectively disambiguate senses can be found in Rayson, Archer, Piao, and McEnery (2004). In principle, a corpus annotated by means of USAS would be of great advantage when it comes to identifying metaphorical usage. As Hardie, Koller, Rayson and Semino (2007) point out, the semantic fields assumed by USAS correspond, approximately, to the "domains" we are familiar with from metaphor theory (WAR, TIME, ACTIONS, STATES etc.). A word such as *campaign* would have multiple semantic tags associated with it, reflecting its varied uses. Hardie *et al.* (2007) discuss the need to implement a "broad-sweep" approach to identifying the relevant semantic tags associated with words like *campaign* where the metaphor researcher might well wish to retrieve both the G3

Wordnet senses of a given term to enable a more efficient annotation by the (human) annotator. For a sophisticated exploitation of WordNet to identify metaphorical senses of words (in the WordNet database, as opposed to searching in a corpus), see Peters and Wilks (2003).

[5] UCREL stands for University Centre for Computer Corpus Research on Language, Lancaster University. More details on USAS can be found at <http://ucrel.lancs.ac.uk/usas/>.

'warfare' semantic category and the X7 'wanting, planning, choosing' category since these two categories represent the source and target domains in a usage like *advertising campaign*. USAS has been implemented in the software suite Wmatrix (RAYSON, 2003; 2007), but to date we still lack publicly available corpora annotated in this way.[6]

Typically, then, exploring metaphorical usage in a corpus will require a good deal of inspection and decision-making by a researcher (see the extensive discussion of issues associated with this task in STEEN, 2007). Boers (1999) stands out, still, as a simply designed but very revealing corpus-based study of metaphor, which involved the manual inspection of all instances of HEALTH metaphors in the editorials of The Economist magazine over a ten-year period. A corpus was constructed, guided by the occurrence of HEALTH metaphors in the editorials over a ten-year period, of about 1,137,000 words from articles accompanying those editorials. Although the resulting corpus may be small by current standards, the task of identifying all HEALTH metaphors in such a corpus (i.e., HEALTH as the target domain rather than the source domain) is not something one would take on lightly. The metaphors identified by Boers include a wide range of forms, as one might expect: *sickly firms*, *diagnosing a shortage*, *the market cure*, *surgery that costs jobs* etc. Some of the expressions were identified as clear instances of a HEALTH metaphor (e.g., *the market cure*), while others were categorized as vague or ambiguous (e.g., *economic remedy*). Boers relied on the Collins Cobuild English Language Dictionary to help make principled decisions about the two categories: when the first (i.e., more frequent) usage in the dictionary entry actually mentioned the domain of physical health, then its figurative use in the corpus was categorized as clear; otherwise the figurative use was categorized as vague or ambiguous. The quantitative data was then presented in two ways – the clear metaphors only and all the metaphors, though the two sets of data were similar.

A somewhat similar, though methodologically more refined, approach to metaphor identification is MIP, i.e., "metaphor identification procedure", as developed by the Pragglejaz Group (2007). MIP seeks to make a clear distinction between metaphorical and non-metaphorical usage and does so in a relatively programmatic way. Thus, a researcher is expected to work through

[6] It has been announced that the International Corpus of English (ICE) will be annotated using USAS and Wmatrix. See <http://ice-corpora.net/ice/index.htm> for details of these corpora and updates on progress.

a series of specified steps to arrive at a decision about the metaphorical status of a word, with the word being the relevant unit of interest. The researcher must determine if the word has a more basic contemporary meaning in other contexts and, if so, whether the word in the use being investigated can be understood in comparison with the more basic meaning.[7] The use of STRUGGLE in a context such as *someone struggled to convince X of Y* is claimed to be a clear case of metaphorical use if the basic meaning of STRUGGLE is taken to be 'use ones physical strength against someone or something'. The use of CONVINCE in the same example is taken to be non-metaphorical since there is no other more basic meaning found in other contexts. As in the case of Boers' approach referred to above, MIP relies, in part, on dictionary entries to guide decisions about basic vs. non-basic meaning. The strength of MIP lies in the explicitness of the procedure and the further testing of the reliability of the procedure (the procedure includes a recommendation that researchers report the statistical reliability of their analyses, e.g., measuring reliability across cases and reliability across analysts) in order to arrive at defensible and replicable results. Even so, the authors in the Pragglejaz Group recognize the challenge of applying their procedure, commenting that "it is not a task that can be accomplished easily or quickly" (Pragglejaz Group 2007, p. 36).

There are, however, various ways in which one might try to reduce the amount of manual inspection associated with exploring source and/or target domain behaviors, recognizing that undertaking a full-blown MIP-type analysis of a large corpus is not feasible (see STEFANOWITSCH, 2006, p. 1-6 for a more extensive review of possible methods):

(a) *Use a small corpus to first identify items of interest* (cf. DEIGNAN 2005, p. 93). Cameron and Deignan (2003) begin with a small corpus of 28,285 words of transcribed talk by primary (elementary) school children to identify, exhaustively, forms and patterns relevant to their interest in metaphor. Their method of identifying metaphorical usage appeals to an earlier insightful discussion of issues surrounding metaphoricity in

[7] An extension of MIP is MIPVU, where VU stands for *Vrije Universiteit.* This extension is described in Steen, Dorst, Herrmann, and Kaal (2010). Among other differences with MIP, MIPVU recognizes a three-way distinction when identifying metaphorically used words: clear metaphor-related words, metaphor-related words that are doubtful, and words that are clearly not related to metaphor. The new "doubtful" category is reminiscent of Boers categories of vague/ambiguous cases of metaphorical usage.

Cameron (1999). Metaphorical usage was identified by Cameron, initially, in cases where there was an incongruity between Topic and Vehicle and where a coherent interpretation of that incongruity is possible. However, Cameron (1999) introduced a number of interesting and subtle qualifications to this general approach. For example, she distinguished "insider" and "outsider" perspectives, depending on whether the interpretation is from inside or outside the shared discourse world of speaker and listener. An example such as "This pillow is my spaceship", as spoken by a three-year old, might be seen as metaphorical for a typical adult hearing such an utterance out of context, but the utterance might be better seen as non-metaphorical from the point of view of the child engaged in creating a particular, imaginative scene and assigning a precise role to the pillow. The results from the smaller and fine-tuned exercise were then used by Cameron and Deignan (2003) as the basis for searching a larger 9 million word corpus (transcribed spoken data from the Bank of English). A variant of this approach uses a sample of texts as a way of first identifying metaphorical uses of interest, e.g., Charteris-Black (2004). When the focus of research is the overarching metaphorical structure of a large piece of discourse (e.g., the rhetoric of an election campaign, the metaphors at work in an advertising campaign, the interplay of metaphorical devices in a work of fiction etc.), then inspecting smaller texts is a natural way to sample the larger discourse, before moving on to other methods of analysis.

(b) *Use a large corpus, but create a fixed set of search terms*. This method is suitable when the research can identify key words in a semantic domain, e.g., domestic animals, weather conditions, modes of transport etc. An electronic thesaurus might be a useful tool to create some sets of related terms. Usually, though, it would be almost impossible to anticipate the full range of expressions, which might instantiate a concept. It would seem very unlikely, for example, that Boers findings about HEALTH, referred to above, could be replicated simply by deciding a priori on a set of forms to investigate. Stefanowitsch (2006) adopts a fairly bold approach in his method for studying metaphorical mappings in the British National Corpus. To explore mappings from the source domain of ANGER, for example, Stefanowitsch identified a representative lexical item associated with this domain, choosing the term with highest frequency from within the set of *anger*, *fury*, *rage*, *wrath* etc. The most frequently occurring term

is, in fact, *anger*, so the form *anger* is the basis for further exploration of mappings from source domain ANGER. While this method may seem somewhat simplistic in its approach, it proves to be surprisingly revealing for the study of emotion metaphors. Another variant of this method is found in Oster (2010) who explores the concept of FEAR, including metaphoric and metonymic uses, in the very large Corpus of Contemporary American English (COCA, <http://corpus.byu.edu/coca/>). She determines, first, lists of collocates of various forms of *fear* and their contexts (maximum 400 hits for each set of results obtained form various search expressions). The identification of metaphorical usage proceeds by applying an adaptation of the morpheme identification procedure proposed by the Pragglejaz Group (2007), working initially with lists of collocates rather than linear text.

(c) *Use a limited number of concordance lines to inspect results*. Deignan (2005, p. 155-157) relies on a sample of 1,000 concordance lines with *cat(s)* as the search term to investigate metaphorical use of these forms. Ideally, such sample concordance lines would be obtained as a random set faithfully representing the range of genres/texts which are of relevance. Stefanowitsch, in the study referred to above, takes a random sample of 1,000 concordance lines to inspect the use of his key emotion terms.

One can arrive at many insights about metaphor from the application of these methods. Deignan (2005) uncovers quite a variety of results which make a very real contribution to the study of metaphor by relying on relatively simple methods like those above. Her discussion of animal metaphors is a good example of just what can be learned by applying corpus-based methods. Deignan (2005, p. 152-157) investigated metaphorical uses of nouns from the source domain ANIMALS (*pig*, *wolf*, *monkey*, *rat*, *horse* etc.). One finding was that simple equational kinds of expressions, like a much celebrated example in the literature (*Richard is a gorilla*), is exceedingly rare in usage. Instead, the animal noun is typically converted to a verb or adjective (*I was horsing around with Katie*; *the mousy little couple*; *she bitched about Dan* etc.). The conversion from noun to verb/adjective in these cases is presumably related to the fact that we employ animal metaphors to conceptualize human behaviors (prototypically expressed through the verb category in English) and, to a lesser extent, attributes (prototypically expressed through the adjective category in English). One would not be able to appreciate these patterns, at least not with any real supporting evidence, without the aid of a corpus. Boers (1999) found

fluctuations in the relative frequency of the health metaphor depending on the month (averaged over the ten-year period), with the highest frequency occurring in the winter months. The explanation that Boers offers is that the winter months are the months when issues of physical health are a relatively salient part of human experience, i.e., the more that health is experienced as an everyday reality, the more likely it is that health will function as a source domain for metaphorical mappings. Boers' study can be seen as further evidence for the experiential grounding of language behavior.

3. Polysemy and synonymy

Cognitive linguistics has been especially interested in exploring semantically based word relations, especially polysemy. Elucidating the nature of the relationships between word-senses, as in polysemy, and the basis for such relationships easily leads to broader discussions about the contexts in which certain words or word-senses appear and the nature of extra-linguistic reality. In their Preface to an important collection of papers on polysemy, Ravin and Leacock articulate two key ideas which emerge from the research represented in their volume:

> […] first, polysemy remains a vexing theoretical problem, leading many researchers to view it as a continuum of words exhibiting more or less polysemy, rather than a strict dichotomy. The second is the increasing realization that context plays a central role in causing polysemy, and therefore should be an integral part of trying to resolve it. Ravin and Leacock (2000, p. v-vi)

There has been some convergence of thinking about how corpora might be enlisted to help integrate context of usage into the analysis of relatedness of word senses (in the case of polysemy) and words (in the case of synonyms).[8] The FrameNet project is one example of this kind of approach in the way it seeks to characterize words and their uses. The FrameNet methodology

[8] WordNet is designed to capture directly facts about polysemy and synonymy and is potentially of great value when integrated with a corpus. See, for example, Davies (2007) who describes an ingenious method of integrating WordNet and the British National Corpus. Davies' proposal facilitates searches based on semantic relationships (e.g., synonyms, hyponyms, hypernyms) and so is potentially useful for tracking down metaphorical usage.

involves, among other things, "examining the kinds of supporting information found in sentences or phrases containing the word in terms of semantic role, phrase type and grammatical function), and building up an understanding of the word and its uses from the results of such inquiry" (FILLMORE; ATKINS, 2000, p. 101). Gries (2006), Gries and Divjak (2009), and Gries and Otani (2010) build upon the same recognition of the role of contextual factors, as evidenced in a corpus, to create their *behavioral profiles* of polysemy and near-synonymy (see below).

A corpus-based approach to analyzing polysemy and near-synonymy would proceed in similar ways, differentiated by the level of analysis: $sense_1$, $sense_2$, $sense_3$ etc. for polysemy, and $word_1$, $word_2$, $word_3$ etc. for near-synonymy. I will illustrate one such approach, guided by both of the key ideas articulated by Ravin and Leacock, as quoted above: the incorporation of a range of contextual factors and the identification of degrees of relatedness between near-synonyms. I will consider the nine slow movement verbs *stagger*, *hobble*, *limp*, *trudge*, *plod*, *amble*, *saunter*, *sidle*, *slink*, already studied by Dąbrowska (2009), but here subjected to a somewhat different analysis. As part of a larger study involving a variety of interesting methods, Dąbrowska had 63 native speakers of English offer definitions of these words and then use the verbs in sentences which illustrate their meanings. While this is not a conventional corpus, which we would normally understand to be a collection of naturally occurring stretches of discourse, the sentences collected in this way can be thought of as an "elicitation corpus" illustrating speakers preferred use of words in context. The sentences were coded for a number of factors: characteristics of the person doing the walking (HUMAN, DRUNK, INJURED etc.), the path (presence of various words/phrases such as *home*, *away*, *in the room*, *into the room*, *from the pub*), the setting (INDOORS, OUTDOORS, COUNTRY) and the manner (CRUTCHES etc.). Dąbrowska's method illustrates perfectly the way in which contextual factors (here, a combination of conceptual/semantic properties and lexical/phrasal forms) can be identified and systematically coded. Twenty sentences for each verb were chosen as the basis for the analysis in Dąbrowska (2009). A subset of the results is given in Table 3. The numbers in this table represent percentages of occurrence of a factor out of the total number of times the verb is used, e.g., 65% of occurrences of *stagger* in the corpus refer to a male walker.

TABLE 3
A subset of six contextual factors relevant to the use of nine verbs, adapted from Dąbrowska (2009: 211, Table 1) and the percentages of occurrence with each verb in a corpus.

	INJURED	MALE	PLURAL	*in the room*	*from the pub*	*home*
stagger	5	65	10	5	40	40
hobble	15	55	0	0	5	0
limp	40	60	0	0	0	10
trudge	0	40	45	0	0	20
plod	0	45	25	0	0	20
amble	0	20	70	0	0	0
saunter	0	60	5	25	0	0
sidle	0	75	5	0	0	0
slink	0	25	5	0	0	0

Dąbrowska judged there to be four clusters of verbs in this group, based on a combination of her own intuitions and an informal similarity judgement study (described in DĄBROWSKA, 2009, p. 210, fn. 6). These are shown in (1).

(1) a. *amble, saunter*
b. *plod, trudge*
c. *sidle, slink*
d. *hobble, limp, stagger*

Relying simply on a visual inspection of the numerical data in Table 3 can quickly leads to quandaries. One can see identical percentages for some sets of verbs: {*saunter*, *sidle*, *slink*}, for example, all show 5% PLURAL in their usage. But what about the pair {*sidle*, *stagger*} where each verb has around 70% MALE factor? Does *sidle* belong more in the {*saunter*, *sidle*, *slink*} group or more in the {*sidle*, *stagger*} group? As we move through more and more data (and remember that Table 3 is only a subset of the complete dataset), "eye-balling" the data to arrive at any satisfying conclusion about the whole dataset becomes impossible. One simply has to turn to statistical methods from the family of multifactorial analysis to make sense of this complexity. One such method is Correspondence Analysis (CA) (BENDIXEN, 2003; GREENACRE, 2007; GLYNN, in press). The overall objective of CA is to represent the maximum possible variance in a plot of few dimensions. The summary statistics given by a CA analysis in the

ca package in R shows that two dimensions explain just 63.5% of the variance. Assuming three dimensions explains a full 86.8% of the variance and so we show the results in the three plots Figures 1-3 showing the interaction of dimensions 1x2, 1x3, and 2x3. "Asymmetric" here means that we can inspect how the verbs lie relative to one another and how the contextual factors are spread out relative to the verbs. The verbs are represented by dots and the factors by triangles. The larger the dot/triangle, the more the contextual factor contributes to the correspondence. In Figure 1, an 'inside' (*in the room*) vs. 'outside' (*from the pub*, *home*) orientation is evident, in Figure 2 an 'injured' vs. 'location' orientation, and in Figure 3 we see a three-way contrast between 'injured', 'inside', and 'outside' factors.

From the plots one can appreciate the closeness of some pairs of verbs by their proximity (though finer details are not so easy to see in the reduced size of the plots shown here), e.g., the pairs {*trudge, plod*}, {*slink*, *slide*}, and {*hobble, limp*} in Figure 1. Notice that these pairs correspond closely to some of the clusters that Dąbrowska had arrived at on the basis of intuition and the judgement task. One also sees some associations between the verbs and the contextual factors by their proximities. In Figure 1, for example, *amble* associates closely with PLURAL (70% PLURAL in Table 3); a number of verbs are close to MALE, most of all *sidle* (75% MALE in Table 3). One can also see that *saunter* is the closest to the "indoors"-orientated contextual factor *in the room*, while *stagger* is the closest to the 'outdoors'-oriented contextual factors *home* and *from the pub*, but these are relatively weaker associations (cf. Table 3, where less than 50% of the use of the verbs have these characteristics).

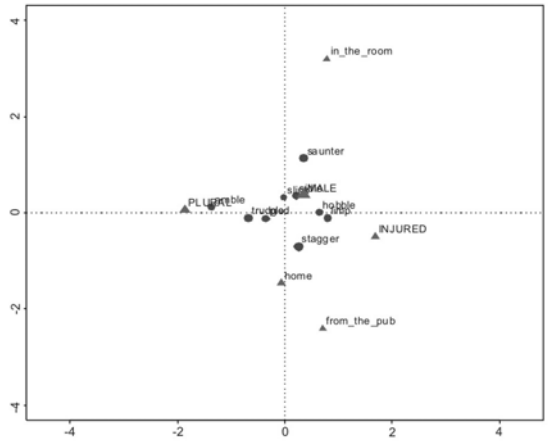

FIGURE 1 - Asymmetric plot of dimensions 1 and 2 from a CA analysis of data in Table 3

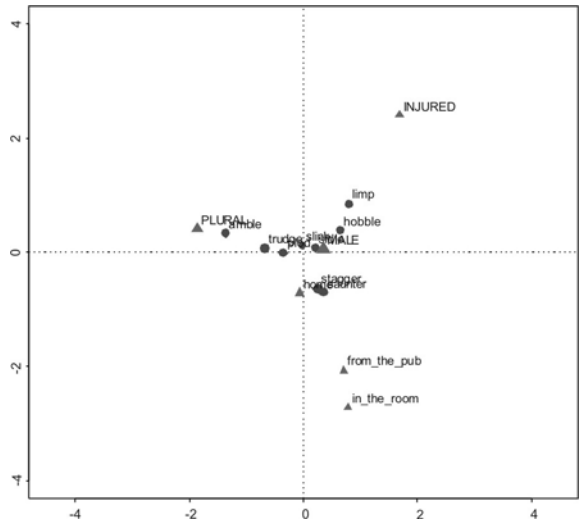

FIGURE 2 - Asymmetric plot of dimensions 1 and 3 from a CA analysis of data in Table 3

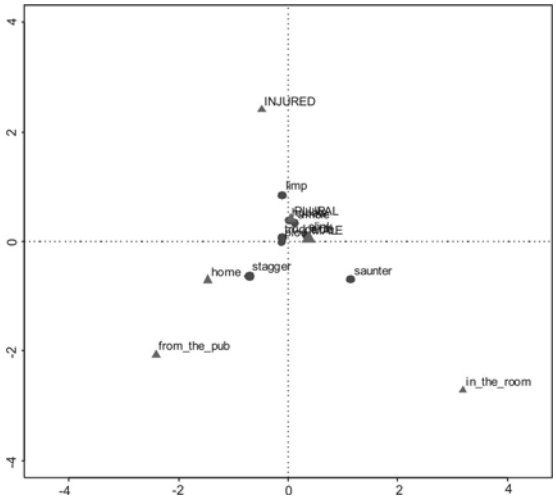

FIGURE 3 - Asymmetric plot of dimensions 2 and 3 from a CA analysis of data in Table 3

Another exploratory method for identifying groupings of verbs is clustering analysis. Clustering analysis subsumes quite a number of techniques, but the basic idea is that numerical data, like that in Table 3, is transformed into a representation by a "distance metric" (a number of options are available), and the transformed data is then the basis for clustering the rows into sub-

groups (again, a number of clustering algorithms are available). Further discussion of these techniques can be found in Gries (2009b, p. 306-319) and Baayen (2008, p. 138-148). Continuing with our sample data in Table 3 for the sake of consistency, we choose from a class of clustering methods called "hierarchical agglomerative clustering" which combines or "agglomerates" the most similar cases (rows in Table 3) into groups and then those groups into larger groups to form a tree-like structure called a "dendogram". In the present case, we choose "Canberra" as the distance metric and a widely used clustering algorithm, "Ward" (ROMESBURG, 2004, p. 101-102, 129-135).[9] The result can be seen in Figure 4. Some groupings emerge as "closer" than others in this tree. The same pairs that we were led to using CA appear here, too, as clusters: {*trudge, plod*}, {*slink, sidle*}, and {*hobble, limp*}. The more vertical height there is between sisters, the less close they are. This means that the pairing of *stagger* with {*hobble*, *limp*} turns out to be a relatively close grouping (consistent with Dąbrowska's fourth grouping above). The dendogram itself does not directly say anything about which contextual factors are contributing most to the groupings visible in the hierarchy. However, by astutely manipulating the input to the clustering analysis (by eliminating one or more columns of data), one can identify certain columns of data as being more or less relevant to some clusterings.

In addition to constructing a dendogram, we should follow up with some tests for the reliablility of clustering produced in a dendogram. Figure 5 includes a probability measure added to each partition of the tree below the root, the "AU" value . The AU value is the abbreviation of "approximately unbiased" probability value (Shimodaira 2004; Suzuki and Shimodaira 2006). One can consider that clusters with high AU values are strongly supported by data. In this case the three groupings that we had identified as being relatively strong by the CA analysis do indeed appear with with high AU values: {*trudge, plod*} = 100% AU, {*slink, sidle*} = 98% AU, {*hobble, limp*} = 93% AU.[10] The dendogram also numbers the clusters on the "edges" in the

[9] The combination of the Canberra distance metric and the Ward clustering algorithm follows Gries' (2006) practice.

[10] The percentages can be expected to vary a little when this procedure is repeated, since the algorithm relies on "multiple bootstrapping", i.e., resampling the original data multiple times, so different sets of sampled data are used as the basis for the calculations. In multiple trials on this data, AU percentages did vary, but only in a small range of a few percentage points.

order in which they are formed (1, 2, 3, etc.). The order here is based on the order of combination of the most similar cases: the lower down on the y-axis where the agglomeration takes place, the sooner the combination takes place in building up the dendogram.

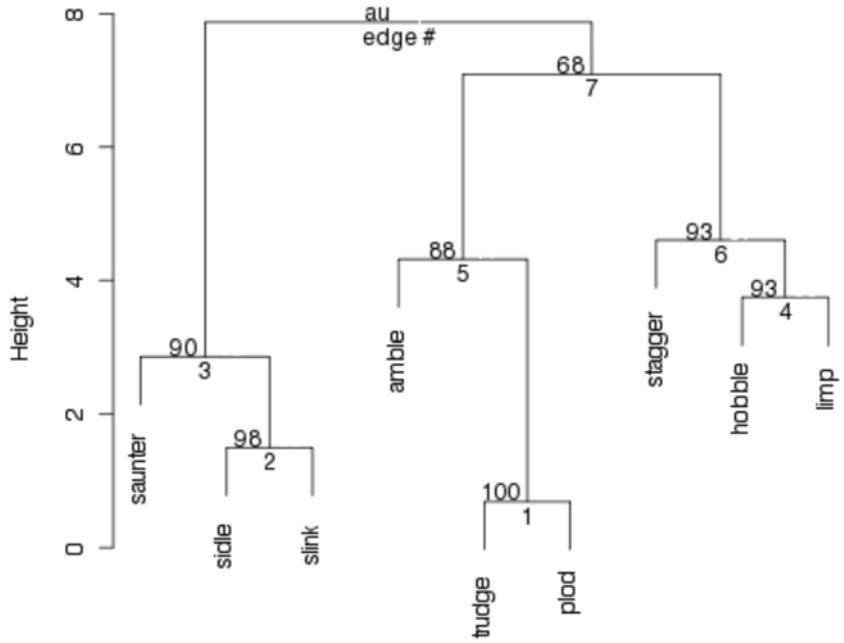

FIGURE 4 - Clustering analysis of data in Table 3

Sometimes, we may have data in a non-numeric format, as in Table 4, taken from Dowbor (ms.). The data in this table are taken from a much larger table in Dowbor (ms.), where Dowbor explored a clustering analysis of the multiple senses of the preposition *over*. Her table was initially constructed in a spreadsheet where she coded each instance of the use of *over* in concordance lines extracted from a corpus – a common way in which such data might be collected. Each sub-sense of over was coded as 'about', 'above', 'across', 'by means of' etc. and these represent the cases to be clustered. Dowbor constructed a number of variables concerning the nature of the verb used with any sub-sense (e.g., dynamicity) and properties of the trajector (TR) and landmark (LM) associated with the use of *over*. Gries and Otani (2010) describe how one might carry out a conversion to numeric-only format, making each feature used in the coding a variable in its own right. The

conversion to a numeric table, along with a number of other attractive features (e.g., addition of probability values to the edges of the dendogram), are contained within the "behavioral profiles" R script (GRIES, 2009a).

TABLE 4
Partial data frame of *over* data, adapted from Dowbor (ms.)

SENSE	Dynamicity	TR	TR_concrete	TR_animate	LM
about	stative	PSYCH	abstract	non-animate	STIMULUS
above	dynamic	THING	concrete	non-animate	PLACE
across	stative	PERSON	concrete	animate	PLACE
across	stative	THING	abstract	non-animate	PLACE
by_means_of	dynamic	COMM	concrete	non-animate	INST
during	dynamic	EVENT	abstract	non-animate	TIME

Polysemy and synonymy can not be properly researched if we insist on only working with discrete yes-no categories. Both types of semantic relationships exhibit gradient properties which must be captured by methods which allow the researcher to appreciate the different degrees to which the usages of a word can be related or the different degrees to which words are similar in their usage. Corpus-based methods which incorporate into their analysis a range of contextual data retrievable from a corpus are particularly suited to revealing this gradience associated with polysemy and synonymy and, moreover, are consistent with the methodological observations made by Ravin and Leacock (2000), cited above. The methods illustrated in this Section presuppose a fair amount of coding of properties of the relevant factors – the data frames which underlie the various statistical analyses above – but the rewards in terms of visualization and conceptualization of the phenomena make this initial outlay of effort well worthwhile.

4. Prototypes

The idea of prototypes is pervasive in cognitive linguistics and, indeed, has been one of the hallmarks of the cognitive linguistic movement. Having one central member of a category, a prototype, is just one way that the members of a category may be organized. Other ways in which a category might be organized include the possibility of multiple "local" prototypes, each with a cluster of other members around it or a whole network of relationships which chain together members of a category. The idea of prototypes stands

in contrast to the view that membership of a category is matter of strict and necessary conditions on all members, without differentiating degrees of membership. Some of the procedures introduced in other sections (including the following section) could be considered as methods to help identify central members of categories. Bybee (2010, p. 76-104) argues for high token frequency within a construction as playing a key role in the formation of central categories, with type frequency and semantics also contributing in important ways to how the central member behaves (e.g., higher type frequency of a word in a construction contributes to the productivity of the construction more than token frequency does).

Stubbs (2001, p. 84-96) outlines a method for systematically studying the central uses of a word and an associated "lexico-grammatical frame" (= construction). The method requires the analyst to work through collocates to identify patterns which are structurally and informationally salient. His method starts with the top 20 collocates in a span of 4 words to the left and 4 words to the right of the word of interest and then examining 20 random concordance lines for each of these 20 collocates. By systematically working through such data, it is possible to obtain a profile of major tendencies within a construction. Stubbs illustrates the method with the verb UNDERGO and successfully identifies key aspects of the construction shown in Table 5.

TABLE 5
Prototypical usage of UNDERGO, adapted from Stubbs (2001, p. 92, Table 4.1)

passive/modal		adjective	abstract noun
forced to		further	medical procedure
required to	UNDERGO	extensive	testing
must etc.		major etc.	change etc.

When working with multivariate data as in the case of Dowbor's data in Table 4 above, it is clear that there are potentially many possible combinations of features. With such data, one would like to know if certain combinations of features stand out as more significant than others in a statistical sense. Conveniently, there is just such a technique, though it is a technique which does not often appear in the handbooks on statistics. This technique is called Hierarchical Configural Frequency Analysis (HCFA) and explanation and illustrations of the method can be found in von Eye (1990), Lautsch and Weber (1995), von Eye and Lautsch (2003), and Gries (2009b,

p. 248-252). HCFA is a procedure which computes the statistical significance of combinations of features that show up in the variables, i.e., the "levels" of the factors in the analysis. Many calculations can be involved when the procedure works through every possible combination and carries out its calculations. In the present case, I used the hcfa_3-2.R script (GRIES, 2004b). Running this interactive script on the TR and LM variables in all of Dowbor's *over* data, the script produces the results in Table 6 (a subset of the full results produced by the script). In this table, we see the combinations of levels of the factors TR and LM where the observed frequency of a configuration is greater than expected in a statistically significant manner, indicated by the three asterisks in the Decision ("Dec") column. Three other statistical results are shown: contribution to Chi-square, probability value of a Holm adjusted probability value (an adjustment applied in order to obtain a more appropriate measure of the contribution of one test when there are multiple tests producing an overall probability), and a measure of pronouncedness "Q" (an effect size, independent of how large or small the data are). The seven combinations of TR and LM values shown are just those seven combinations of these features which yielded results at a very significant level. The example sentences in (2), taken from the ICE-GB corpus which Dowbor used, illustrate these combinations of features. Note that the identification of what we may call "prototypical uses" of *over* in this manner does not rely upon any prior decision about exactly how many sub-senses the preposition *over* has. Table 3 does include coding into sub-senses such as 'above', 'across' etc. under the SENSE variable, but the results in Table 6 were obtained without any reliance upon this particular variable. This is an attractive way to proceed in light of the suspect nature of claims about the number and nature of polysemous sub-senses of a word (cf. the insightful remarks by TAYLOR, 2006, on the problem of polysemy in general and the polysemy of *over* in particular).

TABLE 6
HCFA results showing statistically significant "types" of TR and LM configurations with *over*

TR	LM	Freq	Exp	Chisq	P.adj.Holm	Dec	Q
[AMOUNT]	AMOUNT	60	9.55	266	3.059e-27	***	0.137
EVENT	TIME	56	14.66	116	2.440e-15	***	0.114
THING	PLACE	50	16.67	66	1.220e-09	***	0.093
PSYCH-STATE	STIMULUS	35	3.25	310	1.490e-22	***	0.085
STATE	DEPENDENT	12	0.93	131	5.398e-08	***	0.029
COMMUN	INSTRUMENT	5	0.07	367	1.635e-06	***	0.013
ATTRIBUTE	STANDARD	4	0.05	293	5.107e-05	***	0.01

(2) a. …*over 700 farms* still cannot sell their meat for human consumption [AMOUNT] X AMOUNT

b. the blood pressure <unclear-words> at such a level after repeat measurements *over a considerable period of time* sometimes as long as six months EVENT X TIME

c. a minute on each side on high and then 5 minutes *over a low flame* will do it THING X PLACE

d. In view of the furore *over the transmission of news from the Falklands* PSYCH-STATE X STIMULUS

e. Abortion is the right of a woman *over her own body* STATE X DEPENDENT ENTITY

f. When digital data are transmitted *over a single parallel interface* there is no crosstalk between the codes COMMUNICATION X INSTRUMENT

g. It offered many advantages *over other systems* including rapid action ATTRIBUTE X STANDARD ITEM

Whether one relies on Stubbs' method of identifying the most common usage of a form or statistically more sophisticated methods like HCFA (and collostructional analysis, to be discussed in Section 5), there are very systematic corpus-based procedures which a cognitive linguist may use to identify prototypical usage of a word or pattern. These methods succeed in leading a researcher to choices for prototypes, including, for some kinds of phenomena, multiple prototypes. The methods have the virtue of being

strongly grounded in facts of usage, complementing any other (intuition-based or experimentally based) methods the researcher might employ.

5. Constructions

One area of interest in cognitive linguistics relates to a new understanding we have of the relationship between words and the constructions in which they occur. "Construction" here may include the more traditional chunks of text which correspond to the traditional structuralist view of language consisting of constituents like NP, VP, PP etc. But, with a more open attitude to what is of interest in terms of surrounding context, it can be any one of a number of properties of surrounding context in sentences and utterances which might contribute to a corpus-based analysis of a word in context.

To illustrate some methodological possibilities for the analysis of words in constructions, I will consider the collostructional analysis approach pioneered by Stefanowitsch and Gries (STEFANOWITSCH; GRIES, 2003; GRIES, 2004a). I will illustrate the approach through the examples of two related constructions: EXPERIENCE N and EXPERIENCE *of* N, using Mark Davies' COCA corpus as the basis for the calculations.

In the collostructional analysis methodology, one assesses the statistical significance of the association between a construction and an associated word. Consider, for example, the use of *life* in EXPERIENCE *life* in a corpus (where the small caps indicate the lemma, i.e., all the inflected forms of the verb). There are two pairs of contrasting numbers to be considered in evaluating the significance of the frequency of EXPERIENCE *life*: frequencies of the noun *life* in the EXPERIENCE N construction and the total corpus frequency of the noun *life*, frequencies of the EXPERIENCE N construction and the total corpus frequency of all constructions.[11] Since EXPERIENCE N is an instance of a verb phrase, I take the number of (lexical) verbs in the corpus as an approximation of the total number of relevant constructions of the corpus.[12]

---

[11] There is certainly room for disagreement about what constitutes the "number of constructions" relevant to a problem (cf. SCHMID, 2010 for discussion of this issue).

[12] Bybee (2010, p. 97-101) argues for the prime importance of relative frequency of occurrence within a construction and semantics, rather than Stefanowitsch and Gries' collostructional strength measure. In particular, Bybee objects to the reliance on any assumption that words should be considered as occurring "by chance" in a corpus, an assumption underlying many statistical approaches, including the

Following the procedure described in Stefanowitsch and Gries (2003), we begin by noting the following frequencies in COCA:

(4) From the corpus, we directly obtain:[13]

- the number of EXPERIENCE *life* sequences = 80
- the number of EXPERIENCE + noun sequences = 4,169
- the number of tokens of the noun *life* in the corpus = 293,108
- the number of all lexical verbs in the corpus = 47,560,677

From these we obtain the crucial numbers that constitute the contingency table which is the basis for the statistical calculation. Table 7 presents the 2x2 contingency table (the shaded part in the table) which we will use for the calculation (cf. MANNING; SCHÜTZE, 1999, p. 169-172 for a discussion of the underlying procedure applied to the co-occurrence of words, as opposed to a construction and a word). The numbers in bold in Table 7 are the four numbers from (4), obtained directly from the corpus; other numbers are obtained by subtraction. We then carry out a test of statistical significance, such as the Fisher Exact test, on the numbers in the shaded area and obtain a probability value (6.035451e-18 = 6.035451 with the decimal point moved 18 places to the left). A convenient way to report this value, the "collostructional strength" is to use the negative log to base 10 of this number (=17.22). Intuitively, one can think of this result as follows: there is a total of 47,560,677 datapoints in the total sample; the EXPERIENCE N construction occurs with a probability of 4,169/47,560,677 based on the total number of datapoints; the noun *life* occurs with a probability of 293,108/47,560,677; the joint probability of both *life* and the EXPERIENCE N construction is the product of these two probabilities = 5.402111e-07 which equals 25.69 when applied to

collostructional analysis method. Psycholinguistic evidence can be adduced in support of each of their methods: Bybee cites Bybee and Eddington (2006) in support of her position; Gries, Hampe, and Schönefeld (2005) and Gries, Hampe, and Schönefeld (2010) present evidence for the role of collostructional strength.

[13] One could imagine slightly different ways to obtain the relevant frequencies in COCA. In the present case, I used the following search terms: "[experience].[vv*]" to search for the verb lemma EXPERIENCE; "[nn*]" in the R1 position of "[experience].[vv*]" to search for all tokens of the EXPERIENCE N construction; [vv*] to search for all lexical verb constructions in the corpus.

the total number of datapoints. That is, even if the distributions of *life* and the EXPERIENCE N were completely independent of each other, we would still expect 25.69 occurrences of EXPERIENCE *life*. In fact, *life* occurs 80 times in this construction and this number is more than expected to a statistically significant degree, as shown by the extraordinarily low probability value. So, *life* is "attracted" to the EXPERIENCE N construction.

TABLE 7
Table of frequencies relevant to occurrence of nouns in the EXPERIENCE N construction. Numbers in bold are obtained directly from the corpus. The shaded part is the 2x2 contingency table which is the basis for the statistical calculations

	life	life nouns	Total
EXPERIENCE N	**80**	4089	**4169**
EXPERIENCE N	293028	47263480	47556508
Total	**293108**	47267569	**47560677**

Fortunately, it is not necessary to perform this sequence of steps manually if we use the coll.analysis R script (GRIES, 2004a). This script will calculate collostructional strength for any number of words relevant to a construction. The results from the script for the words ("collexemes") under consideration are shown in Table 8. As can be seen in this table, the script returns the overall word frequency of a collexeme, its observed and expected frequency within the construction, the reliance measure (here called "faithfulness", abbreviated "faith"), the attraction or repulsion of the collexeme to the construction, and the collostructional strength (as computed by the Fisher Exact test in these tables). A collostructional strength >3 is significant at the $p<0.001$ level. Our particular calculations for the *life* collexeme above appear here at rank 8, showing the expected frequency of 25.69 and the collostructional strength of 17.2.

TABLE 8
Collostructional profile of nouns occurring in the EXPERIENCE N construction, based on all genres of COCA, ranked by collostructional strength

RANK	WORDS	WORD. FREQ	OBS. FREQ	EXP. FREQ	FAITH	RELATION	COLL. STRENGTH
1	difficulty	13211	48	1.16	0.0036	attraction	58.7
2	success	49194	52	4.31	0.0011	attraction	36.9
3	difficulties	10123	28	0.89	0.0028	attraction	31.4
4	depression	17654	32	1.55	0.0018	attraction	30.1
5	symptoms	14539	30	1.27	0.0021	attraction	29.9
6	feelings	21766	33	1.91	0.0015	attraction	28.5
7	pain	40596	40	3.56	0.0010	attraction	27.4
8	stress	24492	31	2.15	0.0013	attraction	24.6
9	anxiety	13777	24	1.21	0.0017	attraction	22.4
10	life	293108	80	25.69	0.0003	attraction	17.2
11	problems	100663	43	8.82	0.0004	attraction	15.9
12	wash-out	100	6	0.01	0.0600	attraction	15.3
13	pleasure	17729	20	1.55	0.0011	attraction	15.2
14	discrimination	10200	16	0.89	0.0016	attraction	14.5
15	boredom	2124	10	0.19	0.0047	attraction	13.9
16	discomfort	3146	11	0.28	0.0035	attraction	13.9
17	nausea	1865	9	0.16	0.0048	attraction	12.7
18	frustration	7883	13	0.69	0.0016	attraction	12.2
19	orgasm	1380	8	0.12	0.0058	attraction	12.0
20	burnout	1513	8	0.13	0.0053	attraction	11.7

A number of observations may be made based on the results in Table 8. Notice, for a start, that it is not the case that the ordering mirrors relative frequency of occurrence in the pattern. For example, *life* is in the eighth position even though *life* has the highest frequency of all nouns in table. *Life* occurs relatively often in the whole corpus and so, all else being equal, we would expect more occurrences of this noun in the construction under investigation. Instead of *life*, the collexeme showing the strongest collostructional

strength is *difficulty*. It can be easily seen that the great majority of the top 20 collexemes in Table 10 are nouns which in fact share the same kind of negative nuance that *difficulty* has: *depression*, *pain*, *stress*, *anxiety* etc. And the noun which rises to the first position based on collostructional strength reflects this dominant semantic characteristic. The negative "prosody" evident in Table 10 is not something one could confidently predict from mere reflection on the word. Nor is it a result that is so evident from simply inspecting the most frequent nouns occurring in the EXPERIENCE N construction. The top 20 most frequently nouns in this construction, together with their frequencies, are: *life* 80 *success* 52 *difficulty* 48 *problems* 43, *pain* 40, *music* 34, *feelings* 33, *depression* 32, *things* 32, *stress* 31, *symptoms* 30, *difficulties* 28, *anxiety* 24, *pleasure* 20, *nature* 20, *art* 19, *discrimination* 16, *joy* 15, *frustration* 13. Within this list, almost half the items (*life*, *success*, *music*, *feelings*, *things*, *pleasure*, *nature*, *art*, *joy*) do not show any of the negative prosody so evident in Table 8.

It is also interesting to compare the collostructional profile of the EXPERIENCE N construction with what might appear to be a very similar construction: the EXPERIENCE *of* N construction. Table 9 shows results for a collostructional analysis of this construction, now taking the number of lexical nouns in the corpus as the size of the corpus. In this case, one does not find the same strong tendency towards negative nuances as with the collexemes in Table 8. Instead, the collexemes in the EXPERIENCE *of* N construction include a mixture of negatively nuanced concepts and more abstract, philosophical concepts, e.g., *reality*, *oneness*, *modernity*, *transcendence*, *otherness*. The collostructional profiles in Tables 8 and 9 provide a very convenient way of demonstrating the different types of nouns attracted to what would appear to be, on the surface, similar constructions and lend support to treating the two constructions as objects of study in their own right.

TABLE 9
Collostructional profile of nouns occurring in the EXPERIENCE *of* N construction, based on all genres of COCA, ranked by collostructional strength

RANK	WORDS	WORD. FREQ	OBS. FREQ	EXP. FREQ	FAITH	RELATION	COLL. STRENGTH
1	reality	35688	33	1.24	0.0009	attraction	34.4
2	oneness	355	11	0.01	0.0310	attraction	28.7
3	modernity	2502	15	0.09	0.0060	attraction	28.1
4	life	293108	62	10.20	0.0002	attraction	27.5
5	depression	17654	17	0.61	0.0010	attraction	18.4
6	combat	12963	15	0.45	0.0012	attraction	17.5
7	jealousy	2142	9	0.07	0.0042	attraction	15.7
8	slavery	5994	11	0.21	0.0018	attraction	15.2
9	trauma	6167	11	0.21	0.0018	attraction	15.0
10	pain	40596	18	1.41	0.0004	attraction	13.7
11	stress	24492	15	0.85	0.0006	attraction	13.5
12	giftedness	1401	7	0.05	0.0050	attraction	12.9
13	oppression	3066	8	0.11	0.0026	attraction	12.4
14	suffering	16170	12	0.56	0.0007	attraction	11.9
15	disability	5911	9	0.21	0.0015	attraction	11.8
16	transcendence	967	6	0.03	0.0062	attraction	11.7
17	reading	53958	18	1.88	0.0003	attraction	11.7
18	childbirth	1290	6	0.04	0.0047	attraction	11.0
19	otherness	503	5	0.02	0.0099	attraction	10.9
20	colonialism	1524	6	0.05	0.0039	attraction	10.5

A key point about the interpretation of these tables is that it is the relative order of the nouns which is crucial, more than the precise numerical value. One could make different choices, after all, about how the number of constructions is arrived at which would affect the p-value in the Fisher Exact test. Different choices for the number of constructions, however, would not affect the relative ordering of the degrees of attraction. Indeed, Schmid (2010: 113) opted for a "completely arbitrary number" of 10 million in one such

exercise, while Bybee and Eddington (2006) used the total number of words in their corpus (2 million). Note, too, that one could choose from a number of alternative statistical tests on the 2x2 contingency table, the shaded cells in Table 9. Different tests of significance will yield different numbers as collocational strengths.

Two measures used by Schmid (2010) for describing the relationship between words and constructions are worth mentioning: *attraction* and *reliance*.[14] These measures are calculated as in (5).

(5) a. Attraction = $\frac{\text{frequency of a word in a pattern x 100}}{\text{total frequency of the pattern}}$

b. Reliance = $\frac{\text{frequency of a word in a pattern x 100}}{\text{total frequency of the word in the corpus}}$

Attraction measures the extent to which a particular pattern attracts a word. With respect to the example in Table 7, this would be the equivalent of calculating the frequency of the noun *life* as a percentage of the total (row) frequency of the EXPERIENCE N construction, i.e. (80/4,169)*100 = 1.92%. Reliance measures the extent to which a word appears in one particular pattern versus other patterns. This is the equivalent, in Table 7, of calculating the frequency of the noun *life* as a percentage of the total (column) frequency of *life* in the corpus, i.e. (80/293,108)*100 = 0.03%.

In giving attention above to constructions involving familiar structural units like verb + noun and noun *of* noun, I do not mean to imply that only such units are worthy of interest. On the contrary, many sequences of words which we encounter as n-grams may not have any structure familiar in contemporary linguistic tradition, but are worthy of further study. The very idea that we might learn something of importance from the study of mere sequences of words without giving more prominence to the associated syntactic structure (noun phrases, verb phrases etc.) must seem like an affront to many linguists. It appears to ignore a long tradition within linguistics of

[14] The same measures have also been discussed in Gries, Hampe, and Schönefeld (2005, p. 645-647). Note also that the "faith" score returned in the collostructional analyses shown in Tables 7 and 8 is based on the same proportional calculation as Schmid's reliance measure. The reliance scores are similar to what Janda and Solovyev (2009) incorporate into their constructional profiles.

assigning a hierarchical constituent structure to groupings of words and, in the Chomskyan tradition, seeing rules of language as "structure-dependent". Atkinson, Kilby, and Rocca (1982, p. 149) in a defense of this tradition say: "[…] no serious approach to linguistic analysis looks on sentences merely as sequences of words". I would be inclined to say, rather, that no serious cognitive linguist can afford to ignore the role that sequences of words play, irrespective of what structure one might wish to superimpose on them. Jurafsky, Bell, Gregory, and Raymond (2000), for example, studied different measures of probabilities of occurrence of function words such as *a*, *the*, *in*, *of* etc. and investigated how these measures correlated with phonetic effects such as shortening of the vowel of the function word. They found that a higher conditional probability of the function word given the preceding word predicted vowel shortening in the function word, even in bigrams which are not usually thought of as any kind of structural units such as *them and*, *sometime in*, *where the*, and *fine and.* The study of n-grams along the lines of Jurafsky et al. should interest cognitive linguists, as much as the study of more conventional constructional types.

6. A note on corpora

In the preceding sections, I have sketched out some corpus-based methods which can be profitably utilized by cognitive linguists choosing to work with corpora. These methods assume what must now be considered rather traditional kinds of corpora, by which I mean collections of samples of written usage and transcribed spoken usage from the major languages of the world, relying on an orthographic representation of language. The availability of such corpora and their popularity among usage-based language researchers should not distract researchers from the task of developing and analyzing corpora from other domains of usage. Minority languages and most indigenous languages remain understudied linguistically and underrepresented in available corpora. Even within the major languages of the world, regional and social varieties warrant more attention than they have received in corpus-linguistic circles.[15] Corpus collections of varieties of English such as the

[15] Dirk Geeraerts' research on "cognitive sociolinguistics", as in Geeraerts (1994), incorporates the study of sociolinguistic characteristics of the speakers as part of a larger corpus-based approach, and represents a welcome extension of the usual scope of both corpus linguistics and cognitive linguistics.

International Corpus of English (ICE) go some way towards filling this gap for English, but it is relatively rare for corpus-based studies of English to utilize these (free and downloadable!) corpora.[16] It is more common for researchers of English to seek out corpora which are increasingly large in terms of numbers of words, rather than smaller, specialized ones which are designed to reflect specific kinds of language use.

There is, above all, a pressing need to study the more interactive aspects of spoken language, aspects that can only be investigated with fully multimodal corpora which include and integrate video and audio dimensions. One sometimes encounters in the cognitive linguistic literature references to "situated" language as being a desirable focus, contrasting with de-contextualized snippets of language. Consider, for example, the position articulated in Evans and Green (2006, p. 478) who view "situated instances of languages use" as the basic, raw experience from which speakers build up a mental grammar. From this point of view, the study of "situated instances of language use" is a fundamental aspect of the language experience of speakers, not some peripheral, incidental phenomenon. I endorse this view, but I also believe that situated instances of language use must include a great many more aspects of language use than linguists, including linguists from the Conversation Analysis tradition, are accustomed to thinking about. Certainly, we must go beyond the traditional kinds of data represented in the familiar corpora designed along the lines of BNC, COCA etc. Instead, we must look to data in which hand gestures, head movement, gaze, motion, speed of body movements, facial expressions, and bodily stance are all integrated into the data being investigated (cf. WICHMANN, 2007, p. 82-83, in which the author calls for data from all channels of communication to be included in our corpora). I see Charles Goodwin in publications such as Goodwin (1979; 1980; 1981) as an early pioneer of the approach I am advocating. Another forerunner of such an approach would be Harris (1996; 1998) who has argued forcefully for an agenda for the study of language which situates language well and truly in the context of communication, what Harris calls an "integrationist approach". Thorne and Lantolf (2006) call for a "Linguistics of Communicative Activity", the goal of which is to "disinvent language understood as an object and to reinvent language as *activity*..." (THORNE; LANTOLF, 2006, p. 71, italics original). I believe most linguists, including cognitive linguists and

[16] The homepage of ICE is <http://ice-corpora.net/ice/index.htm>.

corpus linguists, are much more comfortable dealing with language as an object rather than as a process and the "disinvention" that Thorne and Lantolf call for is difficult, even troubling, for many linguists, though it is a change which cognitive linguists should welcome and embrace.[17]

My remarks in this section may be taken as an extended footnote to the whole article. I am most concerned, in this overview of corpora and cognitive linguistics, with assembling analytical methods that cognitive linguists may appeal to in working with corpora as we know them now. However, prevailing ideas about linguistic research and the scope of linguistics necessarily influence and constrain the way such corpora have been designed. Current corpus-based methods may help lead us as researchers to fresh insights language usage, but as long as the corpora themselves reflect only a part of our language behavior in the real world, our methods will still not reveal all that is relevant in language activity.

7. Summary

If cognitive linguistics is to fully develop as a field of linguistics grounded in actual usage of language, then corpora are not just one more type of data to be considered along with other modes of inquiry such as intuition, experimentally based methods etc. More than any other type of linguistic data, corpora represent usage and are therefore, arguably, the most essential kind of data that a usage-based cognitive linguistics should rely on. As mentioned above, not all linguists who identify themselves as cognitive linguists fully subscribe to the view that language usage is a central component of the whole cognitive linguistic enterprise. For those who do see language usage as central, however, corpora must continue to play a critical role in the development of the field.

Along with a focus on corpus data comes a need for new methods to process the data. It is a reflection of the Information Age in which we live that many corpora which are now becoming available are far, far larger than we can easily cope with simply by casting our eyes over concordance lines or columns

[17] MacWhinney's TalkBank and CHILDES projects contain many examples of corpora integrating audio and video. Large-scale examples of such corpora would be the D64 Multimodal Conversational Corpus (OERTEL; CUMMINS; CAMPBELL; EDLUND; WAGNER, 2010) and Deb Roy's Human Speech Genome project (ROY, 2009). The corpora coming out of these two projects capture video and audio in everyday settings in an extremely intensive manner and pave the way for quite exciting new discoveries about situated language use.

of collocates. The magnitude of the data in many corpora is such that linguists must inevitably turn to methods of analysis which will involve some degree of automatic retrieval and analysis. Linguists have no choice but to appeal to techniques of quantitative analysis which have been more familiar and more accepted in some other areas of social science than in linguistics. For this reason, I have focused my attention on methods in the sections above. At the same time, I do not mean to imply that issues about data, as opposed to methods, are relatively unimportant. On the contrary, the collection of multimodal data, incorporating audio and video, and the development of standards for annotating and accessing such data should be high on the agenda for cognitive linguists. Indeed, rethinking language as an activity realized through "situated instances of language use", studied through multimodal corpora, is potentially of far greater consequence to the field than the development of methods for the analysis of corpus data representing "unsituated instances of language use".

References

ATKINSON, M.; KILBY, D.; ROCA, I. *Foundations of general linguistics*. London: George Allen and Unwin, 1982.

BAAYEN, R. H. *Analyzing linguistic data*: a practical introduction to statistics using R. Cambridge: Cambridge University Press, 2008.

BENDIXEN, M. A practical guide to the use of Correspondence Analysis in marketing research. *Marketing Bulleting* 14, Technical Note 2, 2003. Available at: <http://marketing-bulletin.massey.ac.nz/V14/MB_V14_T2_Bendixen.pdf>. Retrieved: April 7, 2011.

BOERS, F. When a bodily source domain becomes prominent: the joy of counting metaphors in the socio-economic domain. In: GIBBS, R. W. JR.; STEEN, G. J. (Ed.). *Metaphor in cognitive linguistics*. Amsterdam / Philadelphia: John Benjamins, 1999.

BYBEE, J. *Language, usage and cognition*. Cambridge: Cambridge University Press, 2010.

BYBEE, J.; EDDINGTON, D. A usage-based approach to Spanish verbs of becoming. *Language*, v. 82, n. 2, p. 323-355, 2006.

CAMERON, L. Identifying and describing metaphor in spoken discourse data. In: CAMERON, L.; LOW, G. (Ed.). *Researching and applying metaphor*. Cambridge: Cambridge University Press, 1999.

CAMERON, L.; DEIGNAN, A. Combining large and small corpora to investigate tuning devices around metaphor in spoken discourse. *Metaphor and Symbol*, v. 18, n. 3, p. 149-160, 2003.

CHARTERIS-BLACK, J. *Corpus approaches to critical metaphor analysis*. New York: Palgrave Macmillan, 2004.

DĄBROWSKA, E. Words as constructions. In: EVANS, V.; POURCEL, S. (Ed.). *New directions in cognitive linguistics*. Amsterdam and Philadelphia: John Benjamins, 2009.

DAVIES, M. Semantically-based queries with a joint BNC/WordNet database. In: FACCHINETTI, R. (Ed.). *Corpus linguistics 25 years on*. Amsterdam and New York: Rodopi, 2007.

DEIGNAN, A. *Metaphor and corpus linguistics*. Amsterdam and Philadelphia: John Benjamins, 2005.

DOWBOR, D. (ms.). *The polysemy of OVER*: a BP and HCFA investigation. University of Alberta.

EVANS, V.; GREEN, M. *Cognitive linguistics*: an introduction. Edinburgh: Edinburgh University Press, 2006.

FASS, D. met*: A method for discriminating metonymy and metaphor by computer. *Computational Linguistics*, v. 17, n. 1, p. 49-90, 1991.

FELLBAUM, C. (Ed.). *WordNet*: an electronic lexical database. Cambridge, MA: MIT Press, 1998.

FILLMORE, C. J.; ATKINS, B.T.S. Describing polysemy: the case of *crawl*. In: RAVIN, Y.; LEACOCK, C. (Ed.). *Polysemy*: linguistic and computational approaches. Oxford: Oxford University Press, 2000.

GEERAERTS, D. Methodology in cognitive linguistics. In: KRISTIANSEN, G.; ACHARD, M.; DIRVEN, R.; Ruiz de MENDOZA IBÁÑEZ; F. J. R. (Ed.). *Cognitive linguistics*: current applications and future perspectives. Berlin and New York: Mouton de Gruyter, 2006.

GEERAERTS, D.; GRONDELAERS, S.; BAKEMA P. *The structure of lexical variation*: meaning, naming, and context. Berlin and New York: Mouton de Gruyter, 1994.

GLYNN, D. Multiple Correspondence Analysis: exploring correlations in multifactorial data. In: GLYNN, D; ROBINSON, J. (Ed.). *Polysemy and synonymy*: corpus methods and applications in cognitive linguistics. Amsterdam and Philadelphia: John Benjamins. (In press).

GOODWIN, C. The interactive construction of a sentence in natural conversation. In: PSATHAS, G. (Ed.). *Everyday language*: studies in ethnomethodology. New York: Irvington, 1979.

GOODWIN, C. Restarts, pauses, and the achievement of mutual gaze at turn-beginning. *Sociological Inquiry*, v. 50, n. 3-4, p. 272-302, 1980.

GOODWIN, C. *Conversational organization*: interaction between speakers and hearers. New York: Academic Press, 1981.

GREENACRE, M. *Correspondence Analysis in practice*. 2. ed. Boca Raton: Chapman and Hall/CRC, 2007.

GRIES, St. Th. *Coll.analysis 3. A program for R for Windows 2.x*, 2004a. url: <http://www.linguistics.ucsb.edu/faculty/stgries/>.

GRIES, St. Th. *HCFA 3.2. A program for R*, 2004b. url: <http://www.linguistics.ucsb.edu/faculty/stgries/>.

GRIES, St. Th. Corpus-based methods and cognitive semantics: the many meanings of *to run*. In: GRIES, S. Th.; STEFANOWITSCH, A. (Ed.). *Corpora in cognitive linguistics*: corpus-based approaches to syntax and lexis. Berlin and New York: Mouton de Gruyter, 2006.

GRIES, St. Th. *BehavioralProfiles 1.01*. A program for R 2.7.1 and higher, 2009a. url: <http://www.linguistics.ucsb.edu/faculty/stgries/>.

GRIES, St. Th. *Statistics for linguistics with R*: a practical introduction. Berlin and New York: Mouton de Gruyter, 2009b.

GRIES, St. Th.; DIVJAK, D. Behavioral profiles: a corpus-based approach to cognitive semantic analysis. In: EVANS, V.; POURCEL, S. (Ed.). *New directions in cognitive linguistics*. Amsterdam and Philadelphia: John Benjamins, 2009.

GRIES, St. Th.; HAMPE, B.; SCHÖNEFELD D. Converging evidence: bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics*, v. 16, n. 4, p. 635-676, 2005.

GRIES, St. Th.; HAMPE, B.; SCHÖNEFELD D. Converging evidence II: more on the association of verbs and constructions. In: RICE, S.; NEWMAN, J. (Eds.), *Empirical and experimental methods in cognitive/functional research*. Stanford, CA: CSLI, 2010.

GRIES, St. Th.; OTANI, N. Behavioral profiles: a corpus-based perspective on synonymy and antonymy. *ICAME Journal*, v. 34, p. 121-150, 2010.

GRIES, St. Th.; STEFANOWITSCH, A. (Eds.), *Corpora in cognitive linguistics*: corpus-based approaches to syntax and lexis. Berlin and New York: Mouton de Gruyter, 2006.

HARDIE, A; KOLLER, V.; RAYSON, P.; SEMINO, E. In: DAVIES, M.; RAYSON, P.; HUNSTON, S.; DANIELSSON, P. (Ed.). Corpus Linguistics Conference, CL2007, *Proceedings*... University of Birmingham, UK, 27-30 July 2007. Available at: <http://ucrel.lancs.ac.uk/publications/CL2007/paper/49_Paper.pdf>. Retrieved: April 7, 2011.

HARRIS, R. *Language and communication*: integrational and segregational approaches. London: Routledge, 1996.

HARRIS, R. *Introduction to integrational linguistics*. Oxford: Elsevier Science, 1998.

HILPERT, M. The German mit-predicative construction. *Constructions and Frames*, v. 1, n. 1, p. 29-55, 2009.

JANDA, L. A.; SOLOVYEV, V. D. What constructional profiles reveal about synonymy: a case study of Russian words for SADNESS and HAPPINESS. *Cognitive Linguistics*, v. 20, n. 2, p. 367-393, 2009.

JURAFSKY, D.; BELL, A.; GREGORY, M.; RAYMOND, W. D. Probabilistic relations between words: evidence from reduction in lexical production. In: BYBEE, J; HOPPER, P. (Ed.). *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins, 2001.

LANDES, S.; LEACOCK, C.; TENGI, R. Building semantic concordances. In: FELLBAUM , C. (Ed.). *WordNet*: an electronic lexical database. Cambridge, MA: MIT Press, 1998.

LAUTSCH, E.; von WEBER, S. *Methoden und Anwendungen der Konfigurationsfrequenzanalyse (KFA)*. Weinheim: Psychologie-Verlags-Union, 1995.

LEWANDOWSKA-TOMASZCZYK, B.; DZIWIREK, K. (Ed.). *Studies in cognitive corpus linguistics*. Frankfurt am Main: Peter Lang, 2009.

MANNING, C. D.; SCHÜTZE., H. *Foundations of statistical natural language processing*. Cambridge, Mass. and London, England: The MIT Press, 1999.

OERTEL, C.; CUMMINS, F.; CAMPBELL, N.; EDLUND, J.; WAGNER, P. D64: A corpus of richly recorded conversational interaction. In: KIPP, M.; MARTIN, J-C.; PAGGIO, P.; HEYLEN, D. (Ed.). LREC 2010 Workshop on multimodal corpora: advances in capturing, coding and analyzing multimodality, *Proceedings*... Valetta, Malta, 2010. p. 27-30. Available at: <http://www.speech.kth.se/prod/publications/files/3433.pdf>. Retrieved: April 7, 2011.

OSTER, U. Using corpus methodology for semantic and pragmatic analysis: what can corpora tell us about the linguistic expression of emotions? *Cognitive Linguistics*, v. 21, n. 4, p. 727-763, 2010.

PETERS, W,; WILKS., Y. Data-driven detection of figurative language use in electronic language resources. *Metaphor and Symbol*, v. 18, n. 3, p. 161-173, 2003.

PHILIP, G. Locating metaphor candidates in specialised corpora using raw frequency and key-word lists. In: MACARTHUR, F.; ONCINS-MARTÍNEZ, J. L.; SÁNCHEZ-GARCÍA, M.; PIQUER-PÍRIZ, A. M. (Ed.). *Metaphor in use*: context, culture, and communication. Amsterdam: John Benjamins. (In press).

PHILIP, G. Metaphorical keyness in specialised corpora. In: BONDI, M.; SCOTT, M. (Ed.). *Keyness in text*. Amsterdam: John Benjamins, 2010.

PRAGGLEJAZ GROUP. MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, v. 22, n. 1, p. 1-39, 2007.

RAVIN, Y.; LEACOK, C. (Ed.). *Polysemy*: theoretical and computational approaches. Oxford: Oxford University Press.

RAYSON, P. *Matrix*: A statistical method and software tool for linguistic analysis through corpus comparison. Ph.D. thesis, Lancaster University, 2003.

RAYSON, P. *Wmatrix*: A web-based corpus processing environment. Computing Department, Lancaster University, 2007. Available at: <http://www.comp.lancs.ac.uk/ucrel/wmatrix/>. Retrieved: April 7, 2011.

RAYSON, P.; ARCHER, D.; PIAO, S. L.; MCENERY, T. The UCREL semantic analysis system. In: Workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks in association with 4th International Conference on Language Resources and Evaluation (LREC 2004), *Proceedings*... Lisbon, Portugal, 2004.

ROMESBURG, H. C. *Cluster analysis for researchers*. North Carolina: Lulu Press, 2004.

ROY, D. *New horizons in the study of child language acquisition*. In: Interspeech 2009, *Proceedings*... Brighton, England. 2009. Available at: <http://www.media.mit. edu/cogmac/publications/Roy_interspeech_keynote.pdf>. Retrieved: April 7, 2011.

SCHMID, H.-J. Does frequency in text instantiate entrenchment in the cognitive system? In: GLYNN D.; FISCHER, K. (Ed.). *Quantitative methods in cognitive semantics*. Berlin and New York: Mouton de Gruyter, 2010.

SHIMODAIRA, H. Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling. *Annals of Statistics*, v. 32, p. 2616-2641, 2004.

STAMOU, S.; ANDRIKOPOULOS, V.; CHRISTODOULAKIS, D. Towards developing a semantically annotated treebank corpus for Greek. In: NIVRE, J.; HINRICHS, E. (Ed.) Second Workshop on Treebanks and Linguistic Theories, *Proceedings*... Växjö: Växjö University Press, 2003.

STEEN, G. J. *Finding metaphor in grammar and usage*. Amsterdam and Philadelphia: John Benjamins, 2007.

STEEN, G. J.; DORST, A. G.; HERRMANN, J. B.; KAAL, A. A. *A method for linguistic metaphor identification*: from MIP to MIPVU. Amsterdam / Philadelphia: John Benjamins, 2010.

STEFANOWITSCH, A. Corpus-based approaches to metaphor and metonymy. In: STEFANOWITSCH, A.; GRIES, St. Th. (Ed.). *Corpus-based approaches to metaphor and metonymy*. Berlin / New York: Mouton de Gruyter, 2006.

STEFANOWITSCH, A.; GRIES, St. Th. Collostructions: investigating the interaction between words and constructions. *International Journal of Corpus Linguistics*, v. 8, n. 2, p. 209-243, 2003.

STEFANOWITSCH, A.; GRIES, St. Th. (Ed.). *Corpus-based approaches to metaphor and metonymy*. Berlin and New York: Mouton de Gruyter, 2006.

STUBBS, M. *Words and phrases*: corpus studies of lexical semantics. Oxford: Blackwell, 2001.

SUZUKI, R.; SHIMODAIRA, H. pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, v. 22, n. 12, p. 1540-1542, 2006.

TAYLOR, J. Polysemy and the lexicon. In: KRISTIANSEN, G.; ACHARD, M.; DIRVEN, R.; de MENDOZA IBÁÑEZ, F. J. R. (Ed.). *Cognitive linguistics*: current applications and future perspectives. Berlin and New York: Mouton de Gruyter, 2006.

THE R PROJECT for Statistical Computing. <http://www.r-project.org/>.

THORNE, S. L.; LANTOLF, J. P. A linguistics of communicative activity. In: MAKON, I., S.; PENNYCOOK, A. (Ed.). *Disinventing and reconstituting languages*. Clevedon: Multilingual Matters, 2006.

VALENZUELA, J. A psycholinguists view on cognitive linguistics: an interview with Ray W. Gibbs. *Annual Review of Cognitive Linguistics*, v. 7, p. 301-317, 2009.

von EYE, A. *Introduction to configural frequency analysis*: the search for types and anti-types in cross-classification. Cambridge: Cambridge University Press, 1990.

von EYE, A.; LAUTSCH, E. Charting the future of configural frequency analysis: the development of a statistical method. [Introduction to a special issue devoted to configural frequency analysis.] *Psychology Science*, v. 45, n. 2, p. 217-222, 2003.

VOSSEN, P. Introduction to EuroWordNet. In: IDE, N.; GREENSTEIN, D.; VOSSEN, P. (Ed.). *Computers and the Humanities*, v. 32, n. 2-3, p. 73-89, 1998. Special issue on EuroWordNet.

WICHMANN, A. Corpora and spoken discourse. In: FACCHINETTI, R. (Ed.), *Corpus linguistics 25 years on*. Amsterdam and New York: Rodopi, 2007.

Appendix. R scripts used for statistical calculations and plots

#Data for reanalysis of Dąbrowska's data in Table 3:
> data # a dataframe constructed in a spreadsheet and imported into R

	INJURED	MALE	PLURAL	in_the_room	from_the_pub	home
stagger	5	65	10	5	40	40
hobble	15	55	0	0	5	0
limp	40	60	0	0	0	10
trudge	0	40	45	0	0	20
plod	0	45	25	0	0	20
amble	0	20	70	0	0	0
saunter	0	60	5	25	0	0
sidle	0	75	5	0	0	0
slink	0	25	5	0	0	0

For Correspondence Analysis of Dąbrowska's data in Figures 1-3:
> library(ca)
> plot(ca(data))
> summary(plot(ca(data))
> plot(ca(x),dim=c(1,2), map="rowprincipal", mass=c(TRUE, TRUE), xlim=c(-3.5, 3.5), ylim=c(-4, 4))¶
> plot(ca(x),dim=c(1,3), map="rowprincipal", mass=c(TRUE, TRUE), xlim=c(-3.5, 3.5), ylim=c(-4, 4))¶?
> plot(ca(x),dim=c(2,3), map="rowprincipal", mass=c(TRUE, TRUE), xlim=c(-3.5, 3.5), ylim=c(-4, 4))¶

For Dendogram of Dąbrowska's data with probability values in Figure 2:
> data.trans = t(data)
> data.trans

	stagger	hobble	limp	trudge	plod	amble	saunter	sidle	slink
INJURED	5	15	40	0	0	0	0	0	0
MALE	65	55	60	40	45	20	60	75	25
PLURAL	10	0	0	45	25	70	5	5	5
in_the_room	5	0	0	0	0	0	25	0	0
from_the_pub	40	5	0	0	0	0	0	0	0
home	40	0	10	20	20	0	0	0	0

> library(pvclust)
> plot(pvclust(data.trans, method.dist = "canberra", method.hclust = "ward"), cex.pv = 1.0, col.pv = c(1, 0, 1)) # to suppress a "BP" probability value and set number font a little higher than the default

#For Hierarchical Configural Frequency analysis in Table 6:
> source("file.path") # run Stefan Gries' hcfa_3-2.R script and follow prompts
OR for similar results
> library(cfa)
> cfa(dataframe)

#For Fisher Exact test on EXPERIENCE *life* in Table 7:
> data<-matrix(c(80, 293028, 4089, 47263480), nrow = 2)
> fisher.probability<-fisher.test(data)$p.value
> fisher.probability
[1] 6.035451e-18
> round(-log10(fisher.probability),2)
[1] 17.22

#For collostructional analysis in Tables 8-9:
> source("file.path") # run Stefan Gries' coll.analysis.R (version 3) script and follow prompts

Recebido em 29/06/2010. Aprovado em 08/05/2011.