Extraction of English Ditransitive Constructions Using Formal Concept Analysis *

Yoichiro Hasebe[a] and Kow Kuroda[b]

[a] Institute for Language and Culture, Doshisha University,
1-3 Tatara-Miyakodani, Kyotanabe-shi, Kyoto 610-0394, Japan
yhasebe@mail.doshisha.ac.jp
[b] National Institute of Information and Communications Technology,
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0289, Japan
kuroda@nict.go.jp

Abstract. This paper proposes a method to extract *constructions* in a formal and mathematically rigid way using the technique of formal concept analysis (FCA). Looking at lemmas of core components of constructions as *objects* and semantic frames of the construction instances as *attributes*, in terms of the FCA, a complete lattice that represents the network structure of constructions can be obtained. We conducted the preliminary experiment of extracting the network of sub-patterns of the English ditransitive construction using a relatively small-sized corpus that was semantically tagged. The result displays the potential capability of this method, both for verifying and substantiating previous theoretical works on construction grammar and for enabling the application of such works to more practical enterprises in the field of natural language processing.

Keywords: construction grammar, formal concept analysis, symbolic thesis, linguistic ontology, ditransitive construction

1 Introduction

In recent linguistic theories based on the symbolic thesis of language, constructions are thought of as fundamental units that cover virtually every level of form and meaning in language, and their importance is greatly emphasized in various areas of syntactic and semantic study (Goldberg, 1995; Östman and Fried, 2008; Fillmore, 2009). There is, however, some obscurity in exactly what meaning is paired with what form of language and, accordingly, what it means when linguists say a construction, such as the ditransitive construction, has various sub-constructions derived from a prototype (cf. Goldberg, 1995). Their argument appears valid, but there is no way to formally substantiate the theoretical entity referred to as a *construction*. Seemingly, it is this lack of a formal definition that impedes the practical application of construction grammar in natural language processing and related fields.

This paper seeks to treat constructions as entities that can be computationally processed, attempting to prove the validity and usefulness of the proposed method by applying it to actual linguistic data. Our method makes it possible to extract the network of constructions from semantically tagged linguistic data without relying on probabilistic analysis. Instead, it utilizes the mathematical technique of formal concept analysis (FCA), originally devised to substantiate *con-*

* The authors are grateful to the anonymous reviewers for their valuable comments and helpful suggestions. The first author was partially supported by the Ministry of Education, Culture, Sports, Science and Technology, Japan, Grant-in-Aid for Young Scientists (B) (No. 21720179).

cepts, by regarding them as being composed of the *extension* and the *intension*.[1] An extension is defined as the set of objects, and an intension as the set of their attributes. It is possible to apply this to language analysis, taking sub-patterns of a construction as objects and the semantic frames represented by actual instances of that pattern as attributes. The network of constructions thus extracted may be suitable for both the verification of previous theoretical works and further computational processing and application.

2 Background

2.1 Constructionist View of Ditransitive Construction

We use the ditransitive construction as a test case for our method, primarily because it is one of the constructions most discussed in the literature and thought of as representative of grammatical constructions in general. Goldberg (1995) illustrates the notable characteristics of this construction using examples such as the following:

(1) (a) *Sally gave her sister a cake.*
(b) *Sally baked her sister a cake.*

The verb *bake* in itself only has the meaning of conducting the process of "baking." The sentence as a whole, however, describes a type of "transfer" of *a cake* from *Sally* to *her sister*. In other words, an additional meaning is added to the literal process of baking, the aspect of meaning reasonably ascribed to the sequence of the indirect object and the direct object. Goldberg considers this to be the effect of the ditransitive construction in English and defines the process of meaning construction involving the ditransitive construction as in Figure 1.

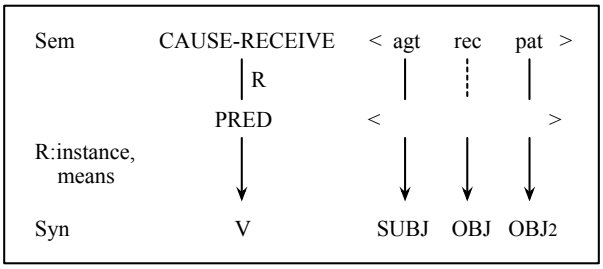

Figure 1: Ditransitive Construction

In the box, the core meaning, or *image schema*, of the ditransitive construction in the semantic pole (Sem) is represented as the function named CAUSE-RECEIVE, which takes three arguments: agent (agt), recipient (rec), and patient (pat). This semantic function CAUSE-RECEIVE also has a variable within it, and the variable is realized, or predicated, by a verb of some particular type that fulfills the restriction stipulated as R on the left-hand side. Finally, this semantic frame, realized by the core image-schema plus a particular verb type with its NP parameters, is paired with the syntactic specification of SUBJ+OBJ_1+OBJ_2, forming the symbolic relationship between the two poles that eventually defines the basic and prototypical structure of the ditransitive construction.

The above CAUSE-RECEIVE schema is considered a conceptual metaphor in the sense of Lakoff and Johnson (1980); as such, seemingly peripheral or even exceptional usages may be approved as far as the divergence from the prototype is motivated by metaphorical mapping between source and target domains. For instance, the prototypical interpretation of the schema "transfer

[1] This means that application of our method does not necessarily require a large data set; a data set of any size could be utilized as the material for analysis. Thus, it is possible, for instance, to analyze data collected from utterances of a child at one particular stage of his/her development using the present method to allow for speculation on the configuration of the child's current linguistic knowledge and capability.

of an item from agent to recipient" can be metaphorically extended to cover "cause-and-result relation between agent and benefactive (recipient)" allowing the occurrence of sentences such as (2). In these examples, the change of spatial position of an object that is originally implied in the construction is abstracted into the change of state in general.

(2) (a) *The medicine brought him relief.*
(b) *The music lent the party a festive air.*

Construction grammar assumes that such metaphorical mapping between domains motivates the production of occasional novel usages that may or may not be later included in the normal use of the language. The new usage will update the whole structure of a particular construction if its frequency is high enough and thus deeply entrenched in a speaker's linguistic knowledge.

2.2 Problems

The constructionist view of language reviewed above is quite reasonable, or at least it seems quite natural to many, supported by the fact that construction grammar has become more popular in recent years. However, for its validity to be verified substantially and computationally, a prerequisite for the theory to be seriously adopted as a means for implementing practical (computational) applications, a few challenges must be overcome.

First, there seems no way to discern the entire semantic range of arguments of a construction. For example, the possible varieties of recipient roles in the ditransitive construction are not specified; it is only clear that the role is that of some kind of recipient of goods or benefits (or possibly harm). Even if metaphorical extension allows virtually infinite varieties of resulting instances, it is necessary to narrow down the possible patterns in order to make that very claim truly plausible.[2]

Second, as is often the case with various sub-fields in cognitive linguistics, construction grammar respects the *gradience* in various aspects of meaning. Accepting gradience as a phenomenon is one thing, and explaining the origin of the gradience is another. It is, therefore, unavoidable for constructionists to attempt to trace the mechanism that creates gradience. To tackle this problem, however, purely speculative investigation is insufficient; it calls for mathematical, computational, and data-driven research programs. If such a mathematical, computational, data-driven approach is not adopted, it cannot be guaranteed that any results or conclusions are not mere speculation reflecting a researcher's personal view on the problem.

With these goals in mind, our proposal for improving the method of investigating constructions is presented in the following sections.

3 Formal Concept Analysis

3.1 Overview of FCA

Formal concept analysis is a technique of analyzing data that enables mathematically rigid definitions of *concepts*. It is based on the lattice theory, utilizing a two-dimensional representation of objects and their attributes in a matrix format called *context* (Ganter and Wille, 1999; Ganter *et al.*, 2005). In FCA terminology, an *object* refers to a thing or event in a domain and an *attribute* to one of the properties that objects could bear in that domain.

More formally, the above definitions are stated as follows: Given a set of objects G and a set of attributes M in a binary relation $I \subseteq G \times M$, a triple (G, M, I) is called a *formal context* and can be represented as a two-dimensional context table, as in the table in Figure 2.

[2] One promising field of study in this respect is that of researchers who are applying advanced statistical techniques onto the analysis of corpus data from a viewpoint compatible with construction grammar (cf. Gries and Stefanowitsch (2006), for general research principles and an introduction to various attempts; Bresnan *et al.* (2007), for a meticulously designed corpus-based study of the English ditransitives).

	bird	mammal	ape	flying	preying	talking
Ostrich	x					
Sparrow	x			x		
Eagle	x			x	x	
Lion		x			x	
Bonobo		x	x			
Human		x	x			x

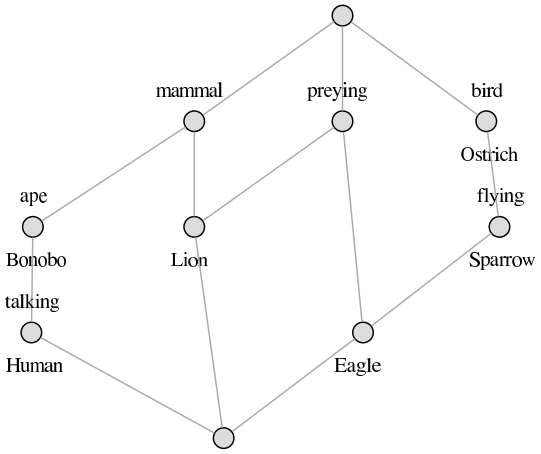

Figure 2: Formal Context (left) and Complete Lattice (right)

The largest set of objects sharing one attribute set is called the *extension*, and the attribute set is called the *intension*. A concept in FCA is defined as the pair of an extension and an intension, and such a concept is specifically called a *formal concept*. For instance, in the table above, the objects in pair {*Sparrow, Eagle*} have attributes {*bird, flying*} as their intension, and the latter attribute pair, in reverse, has those objects as its extension. The double ({*Sparrow, Eagle*}, {*bird, flying*}), therefore, can be considered a formal concept. Formal concept analysis as a mathematical procedure refers to such a process of discerning formal concepts from a given set of objects and attributes.

It is stipulated that two formal concepts (A, B) and (C, D) are considered to have an inheritance relationship (the former is a super concept, while the latter is a sub concept) if, and only if, the following condition is met:

$$(A, B) \leq (C, D) :\Leftrightarrow A \subseteq C (\Leftrightarrow D \subseteq B)$$

The resulting partially ordered set of all the formal concepts derived from a given set of objects and attributes is called a *complete lattice*. Applying formal concept analysis on the context table in Figure 2 (left) and sorting out the inheritance among concepts results in the complete lattice in Figure 2 (right).

Thus, in FCA, concepts are constructed with combinations of two sorts of entities called objects and attributes in a mathematically definable fashion. The resulting complete lattice has a graph structure that enables the relation of concepts to each other, forming a complex, but still highly systematized, network configuration—a favorable condition for application to linguistic data.

3.2 Applying FCA to Linguistic Data

Based both on the theoretical background of construction grammar and on formal concept analysis, we propose a method for extracting the network structure of sub-varieties of a particular linguistic construction—using the ditransitive construction as a sample case.[3] As stated in the previous section, linguistic forms of the construction correspond to objects in terms of FCA, and semantic frames realized in an instance of that construction are considered attributes. In the case of the ditransitive construction, objects and attributes realize as in (3) below:

(3) **objects** *give, send, teach, hand, bake,* etc.

[3] The ditransitive construction in the passive voice is not dealt with here, for in construction grammar, linguistic structures in different voices are considered to form different, though related, constructions.

attributes S:person+O_1:person+O_2:nonhuman-object,
S:person+O_1:abstraction+O_2:abstraction,
S:nonhuman-object+O_1:person+O_2:abstraction, etc.

The lemma forms of predicate verbs are listed as objects. This is motivated by the fact that instances of the ditransitive construction all share the argument structure of S+O_1+O_2; it is enough to register the distinctive part of the form, while abstracting the seemingly irrelevant specifics (here, aspects of verbs such as tense). In the semantic pole of the symbolic structure, combinations of ontological features of argument NPs are listed as possible attributes to the objects. Since there is a characteristic shared by all the semantic frames—namely, the causal relationship among the agent, the recipient, and the theme—this presupposed information does not need to be included in each of the attributes.

From the given data, objects and attributes are detected for each of the instances, respectively producing outcomes that incrementally construct or update the formal context.

(4) **instance** *The party gave the team a chance to get to know each other well.*
object *give*
attributes S:inanimate+O_1:animate+O_2:inanimate,
S:event+O_1:group+O_2:opportunity, etc.

The two semantic frames in (4) are examples that could be constructed differently. Since semantic frames as attributes to objects are not automatically derived from the source, they need to be discovered manually. The extent and granularity of the semantic description should be adjusted according to individual needs. Here, we assign semantic frames on two different levels of linguistic ontology; namely, one on which animate and inanimate objects are distinguished from each other, and the other on which definite categories of thing concepts, such as *event*, *abstraction*, *artifact*, are specified.[4]

4 Experiment

In this section, a preliminary experiment in extraction of the network structure of the English ditransitive construction using data from the semantically tagged corpus SemCor (Mihalcea, 1998) is described. Since language data in SemCor is given semantic information utilizing WordNet's synset system, it maximally reduces the manual (and often error-prone) procedure of discovering and determining sets of semantic frames that are input as attributes for the formal concept analysis.

4.1 Procedure

SemCor is a corpus composed of a subset of the Brown Corpus with syntactic and semantic information added; the senses of content words (nouns and verbs, for instance) are identified with the sense number of the WordNet 3.0 synsets. For instance, the sense of the noun *pencil* in SemCor would be tagged with one of the four synset numbers stipulated in WordNet as different senses of the noun. The differences among synsets are the most apparent if the synsets are traced upward, utilizing the hypernym chain implemented in the system. The hypernyms immediately above the chain for each of the four synsets representing the senses of *pencil* are 1) "writing implement," 2) "graphite," 3) "figure," and 4) "cosmetic." Thus, given the sense number assigned to a given noun instance, its ontological hierarchy can be obtained by tracing the chain up to the point where it reaches the terminal synset (representing "entity").

[4] For an analysis of maximal detail and granularity, we can decompose semantic frames into sets of argument features, produce power sets of these sets, and take each of the members of the resulting power sets as attributes. This will greatly increase the number of the connections in the network—although in many cases to the extent that the complexity of the resulting network repels human scrutiny.

The ontological data gained from this procedure helps in determining the semantic roles of nominals taking part in the construction instances. Even when such detailed ontology data is ready for use, however, using all the information is not a realistic option because of the computational load it produces and over-complexity in the resulting network of constructions. For this reason, we focus only on the two levels of semantic features of NPs that are introduced in (4) in the previous section, the level on which animate and inanimate objects are distinguished from each other, and the one on which definite categories of thing concepts are specified, as illustrated in (5a) and (5b), respectively. These lists are heuristically determined from preceding trials to maximize the coverage of the frames with regard to instances and reduce the accidental overlap of highly distinct concepts in the semantic distribution of a given context.[5]

(5) (a) animate, inanimate
(b) abstraction, act, event, information, nonhuman-object, organization, person

To maximally automatize the process, we also utilized the computational English syntax parser, Link Grammar Parser (Grinberg *et al.*, 1995). Link Grammar Parser detects the dependencies of phrases within a sentence, using a set of linking rules given, and thus identifies its components (a ditransitive verb, an indirect object NP, and a direct object NP, in the present case). Besides, Link Grammar Parser detects the link between a word and another word that is the head of the structure dependent on the former; thus, by combining it with SemCor, which contains WordNet sense numbers assigned to nouns, we can semi-automatically obtain the construction instance's semantic frames based on WordNet sense numbers.[6]

4.2 Result

Out of three folders comprising SemCor, we used the two that contain semantic tags given to nouns, with 186 files of 20,138 sentences in total. Syntactic and semantic data obtained from the procedures using Link Grammar Parser and tracing of WordNet synsets were stored in a database so that these data could be selected, processed, or analyzed as required. For the purpose of building an experimental network structure of the English ditransitive construction, we retrieved 100 instance sentences of the construction.[7] Ditransitive verbs involved in the 100 instances are listed in (6) in the lemma forms with their type frequency.

(6) *give* (57), *offer* (8), *tell* (6), *sell* (6), *bring* (5), *teach* (5) *hand* (3), *cost* (2), *allow* (2), *buy* (1), *get* (1), *pass* (1), *pour* (1), *show* (1), *sing* (1), *write*

Taking these verb lemmas as objects and the semantic frames assigned to their instances as attributes, we applied formal concept analysis, obtaining the complete lattice in Figure 3.[8]

4.3 Discussion

Although the complete lattice in Figure 3 is the result of analysis involving only a small number of instances, it provides several interesting suggestions as to the nature of the ditransitive construction

[5] It was originally intended that the sets of semantic labels be algorithmically determined. The attempt, however, was not very successful mainly because of partial inconsistency of the semantic data gained from WordNet.

[6] Since the semantic frames are constructed with only the roles listed in (5), not all the WordNet sense data are relevant here. The WordNet ontology data are still highly usable since they help in solving the polysemy of the nouns detached from their original context.

[7] We originally tried to gather ditransitive sentences using the data from a parsed subset of Penn Treebank, which also contains data from the Brown Corpus, so that we need not use a computational parsing tool in our study. The sentences collected were too few, however, because the size of the Treebank subset that is also covered by SemCor was quite limited. Thus, we decided to parse SemCor using Link Grammar Parser, first gathering sentences containing a $V+O_1+O_2$ sequence, then (randomly) hand-picking 100 ditransitives from the result.

[8] The complete lattice was graphically rendered with RubyFCA, an open source command-line program developed by the authors of the present paper. (http://kotonoba.net/rubyfca)

in English.[9] First, as pointed out by Goldberg (1995) and other researchers, the subject (S) and the first object (O_1) NPs tend to be realized as animate entities in terms of the token frequency. This can be clearly observed from the fact that the node on right-hand side of the second level from the top (labeled animate+O_1:animate+O_2:inanimate) has far more descendants than its only sister node (inanimate+O_1:animate+O_2:inanimate). Besides, among the concepts that inherit the above semantic frame, those with a nonhuman object as O_2 are quite high in productivity. The attribute label S+person+O_1:person+O_2:nonhuman-object is paired with an object label containing as many as five verb lemmas (*sing, show, pour, pass, hand*). This can be interpreted as showing the prototypicality of this sub-construction.

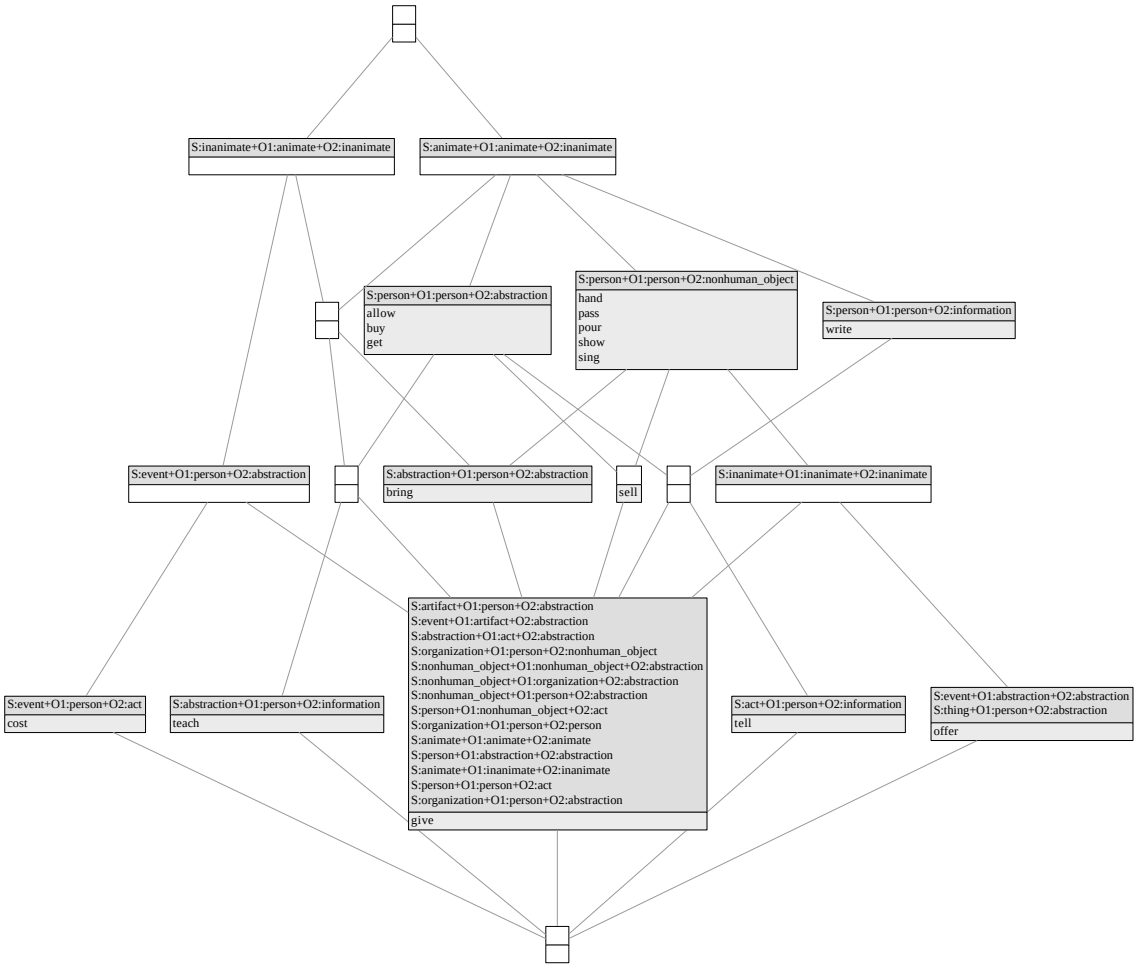

Figure 3: Complete Lattice of English Ditransitive Constructions

Second, we can partially verify the proposition—also maintained by many researchers—that *give* is the default verb of the ditransitive construction in English.[10] The concept node labeled *give* is linked to and inherits from all the nodes on the level immediately above the one in which it is located, and as a result, it has 14 semantic frames out of the total 27. This implies that many other verbs figuring in the ditransitive construction share a basic portion of meaning with *give*.

[9] As the data set is so small, the discussion here is not statistically supported by itself. The argument is, however, highly compatible with Bresnan *et al.* (2007), which offers an analysis of the English ditransitive verbs based on a large-scale statistical survey.

[10] It is, of course, undeniable that the greater variability of *give* relative to other ditransitive verbs results from the high token frequency of *give*'s. Further investigation is required to determine if this truly is the case and to what extent.

There are, on the other hand, concept nodes that have certain semantic frames not shared by *give*; namely, those instantiated by verbs *cost*, *teach*, *offer*, and *tell*. It would be reasonable to say that these verbs have something that is lacking in *give*; the feature that the transfer described is not that of a physical object or a concept that has internal coherency, but rather the transfer of something more abstract such as "act" or "information." The analysis of such peripheral examples of the ditransitive construction utilizing the notion of metaphorical extension should be greatly supported by an observation of the complete lattice presented here.

5 Conclusion

The possible contribution of this study is twofold. First, it conforms extensively to the usage-based model of language, one of the fundamental tenets of cognitive linguistics and construction grammar in particular (Langacker, 1988; Barlow and Kemmer, 2000). Thus, it could naturally support many of the previous works in these fields of linguistics by providing materials for discussion strongly rooted in the actual language data. Second, as the presented method is mathematically and computationally oriented, the materials produced through projects with various theoretical purposes could be utilized to fulfill more practical needs in natural language processing and related areas. Thus, although it lacks multiple tests and is subject to possible modifications, we believe the FCA-based data analysis method will be of importance in both theoretical and pragmatic fields of language studies.

References

Barlow, M. and S. Kemmer, eds. 2000. *Usage Based Models of Language*. Stanford: CSLI.

Bresnan, J., A. Cueni, T. Nikitina, and R. H. Baayen. 2007. Predicting the dative alternation. In G. Bouma, I. Kramer, and J. Zwarts, eds., *Cognitive Foundations of Interpretation*, 69–94. Royan Netherlands Academy of Science.

Fellbaum, C., ed. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Fillmore, C. J., P. Kay, L. A. Michaelis, and I. A. Sag. 2009. *Construction Grammar*. Chicago: University of Chicago Press.

Ganter, B., G. Stumme, and R. Wille, eds. 2005. *Formal Concept Analysis: Foundations and Applications*. Berlin: Springer.

Ganter, B. and R. Wille. 1999. *Formal Concept Analysis: Mathematical Foundations*. Berlin: Springer.

Goldberg, A. E. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press.

Gries, S. T. and A. Stefanowitsch, eds. 2006. *Corpora in Cognitive Linguistics*. Berlin: Mouton de Gruyter.

Grinberg, D., J. Lafferty, and D. Sleator. 1995. A robust parsing algorithm for link grammars. technical report, Carnegie Mellon University Computer Science.

Lakoff, G. and M. Johnson. 1980. *Metaphors We Live By*. Chicago: University of Chicago Press.

Langacker, R. W. 1988. A usage-based model. In B. R. Östyn, ed., *Topics in Cognitive Linguistics*, 127–161. Amsterdam: John Benjamins.

Mihalcea, R. 1998. Semcor: Semantically tagged corpus. Technical report, Southern Methodist University Group of Natural Language Processing.

Östman, J.-O. and M. Fried, eds. 2008. *Construction Grammars: Cognitive Grounding and Theoretical Extensions*. Amsterdam: John Benjamins.