Université Charles de Gaulle – Lille III

Towards a constructional approach to automatic argument structure acquisition: the case of oblique phrases

MÉMOIRE

soutenu le 27 juin 2008

pour l'obtention du

Master de Linguistique de l'Université de Lille III

Master Arts, Lettres, Langues et Communication
Mention Sciences du Langage – Spécialité Linguistique

par

Florent Perek

Composition du jury

	Antonio Balvet	Maître de conférence en linguistique informatique
	Philip Miller	Professeur de linguistique générale et de linguistique anglaise
Directeur :	Maarten Lemmens	Professeur de linguistique et didactique des langues

Laboratoire Savoirs, Textes, Langages — UMR 8163

UMR 8163
CNRS - Universités de Lille 3 et de Lille 1

Mis en page avec la classe thloria.

Contents

Introduction

This study addresses the question of the automatic acquisition of the argument structure of verbs taking a constructional approach to grammar. Its basic goal is to develop a technique to identify argument structure constructions (ASCs) in a corpus, in other words to check whether and if so, how, it would be possible for computers to automatically recognize this type of grammatical constructions on the basis of corpus data. Achieving such a goal would be of major interest in the study of the syntax-semantics interface, since it would provide an objective way to evaluate the claims cognitive linguists. The main benefit of this study would also be to give empirical support to the concept of argument structure constructions and to provide new techniques and tools to researchers in this field.

This work draws on two theoretical frameworks: cognitive linguistics and corpus linguistics. It is an attempt at combining the theoretical aspects of the former (the usage-based model) with the empirical dimensions of the latter. Our model of ASC acquisition relies on two main sources: Goldberg's model of construction grammar and her concept of argument structure constructions and Gries and Stefanowitsch's collostructional analysis, a technique that explicitly aims at reconciliating corpora and a usage-based approach to grammar. More specifically, the technique we develop is based on the assumption that grammar can be derived from corpora, and that a usage-based approach *should* be indeed based on corpora.

In the first chapter, we present the general framework and we detail the main theoretical assumptions and the methodological orientations we base our approach on. The next chapter deals with the potential methodological issues constructions that the notion of contructions poses to corpus analyse (*e.g.* regarding the type of corpus we should chose). We briefly present the Susanne corpus that we have used for our study and the different analyses that we applied to it. In chapter 3, we return in more detail to the model of argument structure construction suggested by Goldberg, on which our model of ASC acquisition is based. We identify the distinctive features of ASCs and how they can be recognised in the corpus. Chapter 4 presents the empirical core of this study, the design and critical evaluation of several indices that were used on the Susanne corpus in order to provide evidence for ASCs.

The results basically show that while our quantitative approach is in line of what is to be expected, it is nevertheless insufficient, because of the size of our corpus and, more importantly, because of the limits of an account exclusively based on formal cues. In our conclusion, we present perspectives that could possibly overcome those problems.

Chapter 1

Motivation and theoretical background

1.1 Cognitive approaches to language

1.1.1 Fundamental claims of Cognitive Linguistics

Throughout the second half of the twentieth century, trends in linguistic theory have been dominated by formalist approaches, which share the idea that linguistic structures, and in particular syntactic structures, can be regarded as manipulation of forms devoid of any meaning[1]. The success of Chomsky's *Syntactic Structures* and subsequent publications (Chomsky 1957; 1965a) established the generative paradigm as a reference framework for linguistics, and the successive versions of generative grammar elicited more and more interest in the linguistic community. Generativists basically seek to posit explicit devices capable of producing all and only the sentences of a given language.

In the seventies and early eighties, more and more scholars, especially in the eastern United States, were dissatisfied with the explanatory power of formalist approaches and began to seek new theoretical orientations, under the influence of new theories and findings in the other cognitive sciences, particularly cognitive psychology. This shift in paradigm gave rise to a new linguistic school of thought, commonly known as Cognitive Linguistics, that aims at "investigating the relationship between human language, the mind and social-physical experience" (Evans et al. 2007).

Cognitive approaches to language are motivated by several guiding principles, which Evans et al. (2007) sum up in two main commitments: the cognitive commitment and the generalization commitment. The cognitive commitment refers to the explanation and characterization of the principles of language in accordance with what is known about the human mind (Lakoff 1990). It thus relies on the findings of other cognitive sciences, such as psychology, artificial intelligence, cognitive neuroscience and philosophy. This principle poses that language is deeply integrated in cognition. The model of language should reflect what is known about the mind rather than seek economy of representation or a computationally efficient formalization (Croft 1998). For example, it is assumed in cognitive psychology that categorization is based on prototypes (Rosch 1973; 1978); the prototype theory has been successfully applied to word meaning as well: color names (Berlin and Kay 1969), action verbs (Pulman 1983), adjectives (Dirven and Taylor 1988), prepositions (Vandeloise 1986) as well as demonstratives (Fillmore 1982) were evidenced to display prototype effects (see Lakoff (1987) and Kleiber (1990) for a thorough review of the prototype theory and its applications to linguistic semantics).

The generalization commitment corresponds to the characterization of general principles that apply to all aspects of language, so as to reach the broadest generalization possible inside the language model. By

[1] Those approaches also often make use of formalisms inspired by computer science, mathematics and logics to obtain a precise formulation of linguistic facts; but the term *formalist* does not refer to the attempt at formalising linguistic insights, which is a much needed and laudable enterprise.

contrast, formal approaches usually scatter the language faculty into separate modules, such as morphology, phonology, syntax, semantics, etc. In the formal model, those 'modules' are structured by different principles and operate on different primitives. For example, a syntactic theory accounts for how words are combined to form complex expressions, a theory of phonology deals with how sounds are structured into phonemic patterns, and so on. Cognitive linguistics adopts a different stance in seeking to uncover "how the various aspects of linguistic knowledge emerge from a common set of human cognitive abilities upon which they draw" (Evans et al. 2007:4). For example, we just mentionned above that the prototype theory was first applied to word meaning; but it has since then been successfully extended to various domains of linguistics such as morphology (Taylor 1995) and phonology (Jaeger and Ohala 1984). The generalization commitment has the effect of giving a unified and exhaustive account of language facts and of abolishing the boundaries between the various traditional domains of linguistics.

1.1.2 The usage-based approach

Another common assumption in cognitive linguistics is the idea that language is usage-based (Langacker 1987; 2000, Bybee 1995; 2003; 2006, Bybee and Thompson 1997, Barlow and Kemmer 2000). The language ability is anchored to its actual usage and emerges from the linguistic stimuli that the speakers perceive during what Langacker (1987) terms *usage events*. Linguistic knowledge is strongly tied to the speaker's *experience*; as Bybee (2006) phrases it, "a usage-based view takes grammar to be the cognitive organization of one's experience with language" (p. 711). In arguing for a usage-based approach, Langacker attempts at explaining why there has been so much emphasis on the description of rules in generative grammar:

> Language is a mixture of regularity and idiosyncrasy. By training and inclination, linguists are better equipped to deal with the former than the latter, with the consequence that far more effort goes into the formulation of general rules than into the patient elucidation of their limitations [...]. The notion of a usage-based model represents an attempt to redress this imbalance, and to overcome the problems it engenders. (1987:411)

Adopting a usage-based approach implies a number of fundamental ramifications. One is that the mental grammar of the speaker results from the abstraction over situated language instances: "grammatical rules are simply a schematization of particular expressions" (Langacker 1991:46). There is a continuum between so-called rules (patterns of usage) and stored exemplars. Units of all degrees of schematicity are assumed to coexist in the grammar: "specific expressions are capable of achieving the status of conventional units even when their formation is perfectly regular" (*ibid.*), thus avoiding what Langacker terms the "rule-list fallacy".

Another important consequence of the usage-based approach is that it assumes no principled distinction between knowledge of language and language use: "knowledge of language *is* knowledge of how language is used", (Evans et al. 2007). They are seen as intertwined phenomena. The usage-based model actually denies the basic chomskyan distinction between competence and performance.

1.1.3 Construction Grammars

Construction grammar is more a family of similar frameworks than a single theory. It is closely related to cognitive linguistics and is the framework par excellence for cognitive analyses of grammar. Each version of construction grammar has its own specificities; however, all constructional approaches to grammar share the same central features and commitments.

To start with, grammar is viewed as an inventory of linguistic units, called constructions. A construction is treated like a sign in the Saussurean sense, *i.e.* form-meaning pairs. Form and content are symbolically linked. As in Langacker's (1987, 1991) Cognitive Grammar, everything in a construction grammar is a construction: from morphemes and lexical words to syntactic templates and idioms; in fact, as in most (if not any) cognitive approaches to grammar, the traditional distinctions between those linguistic elements is thought of as varying positions on a schematicity-idiomaticity cline. Table 1.1.3

sums up some examples.

Construction type	Traditional name	Examples
Simple, not schematic	morpheme	-ing
Simple, not schematic	simple word	book, father
Complex, not schematic	complex word	worker, bookshop
Simple, schematic	inflection rule	plural [-s]
Complex, schematic	syntax	passive
Complex, very schematic	syntax	ditransitive
Complex, partially schematic	(formal) idiom	to kick the bucket the X-er the Y-er X let alone Y (Fillmore et al. 1988) what's X doing Y (Kay and Fillmore 1999)
Complex, not schematic	idiom	A rubber cheque

Construction grammar rejects modularity: it is a monostratal theory. It is also non derivational. All aspects of language, that are traditionally dealt with by different modules (morphology, syntax, phonology, and even pragmatics) are here accommodated in the same framework. Unlike generative grammar, construction grammars fulfill the generalization commitment of cognitive linguistics in positing no autonomy of syntax from semantics, and no clear cut between syntax and lexicon but rather a continuum.

The most striking difference between construction grammar and generative grammar is that the former is a top-down approach while the latter is a bottom-up approach; or in other words, generative grammar (not unlike most related approaches) is a reductionist account, while construction grammar is a holistic account. Generative grammar seeks to uncover some set of systematic rules to predict the meaning of a whole expression from the meaning of its parts; when some expression causes problem to this generality and systematicity, there is a tendency to call the expression involved *idiomatic* and treat it as an exception explicitly declared in the lexicon. On the other hand, construction grammar is a more flexible approach that takes into account the meaning of full expressions or templates. It seeks a compromise between predictability and generalization: the meaning of expressions that do not follow a "normal" syntax is actually predictable to some extent.

The most classical construction grammar is that of Charles J. Fillmore and Paul Kay (Fillmore et al. 1988, Kay and Fillmore 1999, Kay 2002); it looks more closely at the formal aspects of constructions by providing a unification-based framework, fairly much in the same way as Head-driven Phrase Structure Grammar (Pollard and Sag 1987; 1994), with which it shares many theoretical assumptions. Another classical trend in construction grammar is the Lakovian/Goldbergian style (Lakoff 1977; 1987, Goldberg 1995; 2005); while based on the same assumptions as Fillmore's, it stresses less the formal features of constructions and focuses instead on the relations between constructions. Langacker's Cognitive Grammar (Langacker 1987; 1991), as we said, is very similar to construction grammar: it indeed shares the same basic assumptions while adopting a slightly different terminology. Cognitive Grammar also focuses on the cognitive grounding of linguistic explanation, thus placing broad-range cognitive process at the heart of the grammar. Croft's radical construction grammar (Croft 2001; 2004) is designed to account for cross-linguistic factors and thus is especially suited for typological studies. It is much more radically constructionist (hence the name), in the sense that constructions are not derived from their parts, but rather that the parts are derived from the constructions they appear in. Croft denies the existence of universal categories and grammatical relations, and argues that they are not only language-specific, but even construction-specific.

1.2 Current issues about argument structure

Against the background of these theoretical assumptions, the present study considers one particular phenomenon that has received some attention within Construction Grammar (especially in Goldberg's

(1995, 2005) work), *viz.* that of argument structure constructions.

1.2.1 The linking problem

One of the major issues in linguistics is the semantic compositionality of complex expressions, of which argument structure constructions are one particular instance. The issue is how the meaning of complex expressions such as *red ball* or *Jim walks* comes about from the combination of the smaller units that the phrase is composed of. This is often referred to as the linking problem. It is so-called, because it addresses the question of how the meaning of the smallest units are linked to the resulting composite semantic structure. The meaning of clauses is commonly agreed to depend directly on the verb and to involve the linking of the components of some syntactic pattern the verb appears in, to a resulting complex semantic structure that usually depicts a state, an action or more generally, an event (in the sense of Vendler 1967). The various participants involved in such an event are called arguments, and the different argument configurations verbs can appear in are called argument structures. For example, the semantic structure of a clause such as *Jim walks* is an event where the argument *Jim*, that most grammars, if not all, would call the subject, is interpreted as the walker in a walking event expressed by the verb.

Of course, the linking patterns we described for the examples *red ball* and *Jim walks* are valid not only for those examples, but can be generalized in order to derive systematic linking principles. One concern of linguistics is to seek a proper level of generalization in order to account for such expressions.

The following of this section is meant to present a constructional approach to the linking problem applied to argument structure. First, we will present previous approaches in 1.2.2. Subsequently we will introduce argument structure constructions in 1.2.3.

1.2.2 Previous approaches

Many previous accounts of argument structure take a lexical point of view, in the sense that they consider the lexical semantics of the verb as the basis of the explanation. Verbs are often assumed to cast several syntactic configurations, in the generative tradition called *subcategorization frames*. The basic hypothesis underlying many of these accounts is that the syntactic subcategorization frames of a verb stem can be predicted from the verb's meaning.

The semantic structure (often also called "semantic representation") of a verb is a simplified semantic form combining basic operators and referent variables. Consider for example the following semantic structure for verbs of putting quoted from Levin and Rappaport (1995:24):

(1) a. Verb of putting: [x CAUSE [y BECOME P$_{loc}$ z]]
    b. Gloss: x cause y to be in location z

Each syntactic configuration is actually related to a distinct semantic representation. In doing so, this account succeeds in capturing the insight that changes in complement configuration are semantic. The semantic difference between patterns is usually subtle and sometimes has no consequence on the truth conditions, but may rather express a different perspective on the situation being described. For example, the causative and non-causative uses of *break* are represented by Levin and Rappaport (1995:23) and Levin (1999) by the semantic structures in (2a) and (3a) respectively (glossed in (2b) and (3b)) :

(2) a. [y BECOME *BROKEN*]
    b. Gloss: x becomes broken

(3) a. [[x ACT$_{<MANNER>}$] CAUSE [BECOME [y <*BROKEN*>]]]
    b. Gloss: x acting in some manner causes y to become broken

The semantic structure is mapped onto its syntactic realization via general linking rules that capture regularities in syntax to be as broad a generalization as possible. The different semantic structures of a given verb stem are not supposed to be arbitrarily posited but are related by creative lexical rules that take a verb with some particular semantics as input and yield a verb with a slightly different meaning. From this assumption follows the widespread theoretical choice of positing several lexical entries for each verb.

1.2.3 Argument Structure Constructions

The specificities of the constructional approach to grammar, described in 1.1.3 allows other possibilities to account for argument structure. As said earlier, previous approaches are essentially lexicalist and reductionist which means that they seek to analyse the meaning of utterances as some combination of the meaning of atomic units (*i.e.* words and morphemes). The lexico-semantic rules account for the existence of independent constructions. A constructional approach, on the other hand, allows the existence of any kind of syntactic pattern of any level of schematicity. Argument structure constructions can thus be posited as generalizations of the argument structure of verbs over individual instances.

These constructions contribute their own semantics and can be extended to various verbs as long as semantic compatibility is respected. For example, the sentence *Sally baked Bill a cake* implies a transfer of the cake to the recipient Bill; this implication is contributed by the ditransitive argument structure construction only, since the verb *bake* does not plausibly encode a transfer. A similar sentence with the verb *give* would feature the same construction, but the recipient argument would then be contributed by the verb.

By explicitly distinguishing the verbs from the complementation patterns, this account allows a focus on constructional meaning, the relations between verbs and construction, and the relations between constructions (*cf.* Goldberg 1995:chapter 1). In addition, it features the advantage of reducing polysemy and avoiding implausible verb senses. For the lexical semantic approach to account for such examples as (4) (taken from Goldberg 1995),

(4) Sam sneezed the napkin off the table.

it would need to posit an additional meaning of the verb *sneeze*, which would basically be 'X causes Y to move to Z by sneezing'. Such a meaning is indeed counter-intuitive and implausible.

Goldberg's constructional account thus gives more weight to the construction as a linguistic unit with its proper semantic and syntactic properties. One of the issues that needs to be studied then is the interaction between verbs and constructions, which is one of the issues the present study adresses, on the basis of a quantitative and qualitative corpus study. As the next section will show, corpus linguistics is well suited to empirically verify the usage-based approach that cognitive-functional linguistics advocates.

1.3 Cognitive corpus linguistics

1.3.1 Corpus linguistics

Corpus linguistics is the study of language as found in naturally occurring samples, or "real world" text, usually provided by vast repositories, *i.e.* corpora. Corpus linguistics rejects the susceptibly abusive use of introspection in linguistic research and insists on the reliance on empirical data as provided by corpus data in linguistic studies, a position that has often been criticized by Chomsky and generative linguists.

As a matter of fact, many early linguistic studies were empiricist, and thus made use of much corpus data (see McEnery and Wilson (2001:2–4) for a concise review of early corpus linguistics). This is particularly true of distributionalism whose aim was in the first place the study of Amerindian languages, precisely through the systematic investigation of naturally occuring language data gathered in a corpus. Most examples of pre-chomskyan corpus studies indeed come from American linguistics, though not exclusively; the pre-sausurrian case of Käding (1897) is particularly noteworthy. After all, such a tendency should not be surprising if we consider that linguistics of the twentieth century as we know it sprouted under the influence of more traditional disciplines such as philology and hermeneutics, which were dedicated to the collection and analysis of ancient texts.

Chomsky's book *Syntactic Structures* (Chomsky 1957) marked a turning point. In his book, Chomsky pointed out several major flaws of corpus vis-à-vis the theoretical positions taken. Chomsky's proposals were intended to lead linguistic research from empiricism to rationalism, and some of his claims argued strongly againt the use of corpora. First, one of the cornerstones of Chomsky's work is the claim that linguists must seek to model linguistic competence. However, since corpora are by their very nature performance data, they should be considered irrelevant to the comprehension of linguistic facts. According to Chomsky, naturally occuring data is a poor mirror of our competence. Corpora offer no way to figure out what features should be accounted by performance and to what extent their data are representative of linguistic competence. Indeed, as corpora are intrinsically finite and language is infinite, it is hard to argue in any way that the former are adequately representative of the latter, as the following quotation by Chomsky reported in (McEnery and Wilson 2001:8) points out:

> First, it is obvious that the set of grammatical sentences cannot be identfied with any particular corpus of utterances obtained by the linguist in his field work. Any grammar of a language will project the finite and somewhat accidental corpus of observed utterances to a set of (presumably finite) of grammatical utterances. In this respect, a grammar mirrors the behaviour of the speaker who, on the basis of a finite and accidental experience with language, can produce or understand an indefinite number of sentences.

Second, corpora were criticized for their inability to present negative evidence. While generative linguists usually make extensive use of grammaticality judgments, contrasting grammatical from ungrammatical examples in order to infer general rules, corpora are indeed incapable of providing such data: by definition, there is no corpus of ungrammatical sentences.

These claims by Chomsky and generative linguists, which were accurate to some extent, led many linguists to favour introspective judgments, sometimes exclusively, and disregard corpora in linguistic research. This methodological turn deeply undermined the development of corpus. However, not all linguists actually abandoned the corpus trail.

First, in the fields of phonetics and language acquisition, corpora were and still are today the primary source of evidence anyway, since introspective judgements are not available and are just irrelevant. Chomsky himself (1965b) acknowledged that the rejection of performance data as a source of evidence was inappropriate for language acquisition studies.

Some researchers also argued that some of Chomsky's objections towards corpora were slightly overexagerrated. Moreover, as the proponents of corpus methods argued, corpora offer benefits that introspective data, despite its accessibility and flexibility, would never be capable of. Naturally occuring data has the advantage of being observable and verifiable, contrary to the thought processes involved in introspective judgments. That is why Leech (1992) argued that using a corpus is a more powerful methodology from a scientific point of view. The analysis of such data is usually more objective than introspective judgments, which is why corpus-based studies remained valuable and influential. Moreover, quantitative data such as frequencies can only be provided by corpora.

Throughout the 60s and 70s, the corpus methodology survived, though remaining somehow a minoritory. British linguists like Randolph Quirk, Sidney Greenbaum, Jan Svartvik and Geoffrey Leech collected and used (now famous) corpora to study the English language and design grammar books (Quirk et al. 1972; 1985). Some others, like John Sinclair and Michael Stubbs, published influential work in the study of collocations (Sinclair 1991; 1996, Stubbs 1995; 2001).

In the beginning of the eighties, the computerization of corpora was a breakthrough for the field. It allowed huge amounts of data to be easily collected, archived and processed, which opened the way for the development of corpus methodology. As computers were getting more and more common, it boosted the development of corpora and tools to harness them: search for specific patterns, retrieve and sort data, compute calculations. Corpora grew bigger and thus more representative, which somehow reduced the relevance of the argument about the finiteness of corpora. Nowadays, the word *corpus* should actually be understood as *machine-readable corpus.*

The use of corpora in linguistics is now considered by most researchers as complementary to introspection: even while corpora corpora might not always be exhaustive, they often reveal linguistic facts that researchers would never have thought of on the basis of their intuition. Such a position is defended by Fillmore (1991:35):

> I have two main observations to make. The first is that I don't think there can be any corpora, however large, that contain information about all of the areas of English lexicon and grammar that I want to explore; all that I have seen are inadequate. The second observation is that every corpus that I've had a chance to examine, however small, has taught me facts that I couldn't imagine finding out about in any other way.

But as we will see, some very recent proposals are meant to cast off the current limits of corpus-based analysis of language with new methods that conciliate corpus and grammar.

1.3.2 Corpora and the usage-based model

Let us now turn back to cognitive linguistics. What motivates the use of corpora in a cognitive approach to language? How can repositories of textual data be of any use to understand and explain language through cogntive processes? Corpus data is actually very well-suited for cognitive linguistics, because of two reasons which have to do with the usage-based model.

First, the usage-based model reconciliates linguistics with the study of performance. Cognitive linguists deny the generativist distinction between performance and competence, arguing instead that grammar emerges as a result of language usage. Since cognitive linguistics commits itself to the conjoined study of performance and competence and not only the latter, the chomskyan objection about corpus performance data being irrelevant for linguistic analysis is invalid. In this view, corpus data is even more valuable than introspective judgements, because it is actual language use and not artificial data tailored to the purpose of the theory.

Corpora indeed allow one to uncover one aspect of language that is all too often neglected: the actual usage of language. This is why corpora are a valuable source of data for cognitive linguists. According to the usage-based model theory, language acquisition and organization of the language are supposed to be highly determined by its usage, taken as an input stimulus. So, for cognitive linguists, corpora are a first approach to the study of this input that speakers are exposed to and that allows them to learn a language and develop their language ability. Of course, the type of input children are actually exposed to is expected to be very different from the type of language data that can usually be found in corpora, especially if we argue that parent-child exchange has a preponderant role. Corpus data has to be considered as a mere approximation of it. But if we take corpora as a source of naturally occuring data, we can assume that we can derive a grammar model from them, just as language users do when they learn and develop their mother language (or other languages they might learn later on their life).

1.3.3 Corpora and constructions

Long before cognivists got interested in corpora, corpus linguistics pioneers such as the British linguists John Sinclair and Michael Stubbs, obtained enlightening results in the study of collocations (Sinclair 1991; 1996, Stubbs 1995; 2001). However, those trials appeal to a purely linear and co-textual approach that is blind to the grammatical constructions (in the traditional sense) in which those collocations are realized.

Lemmens (1998) is one of the first cognitive studies based on a huge amount of corpus data integrating both lexical semantics and grammatical constructions. It addresses the complexity of the semantic interaction between verbs and constructions. This study revealed the limitations of both the purely constructional approach (Goldberg 1995) and the lexical approach (Levin and Rappaport 1995), but as it needed a manual constructional analysis, its scope was still fairly limited. Such manual processing is indeed long and tedious; therefore samples have to be limited in size, which can reduce the quality and relevance of results, particularly if statistical data are involved. Moreover, it also raises the problem that such studies inherently rely on the researcher's intuition and judgement about the identification of constructions. Considering a more objective way to recognize constructions is thus an issue of primary importance.

Gries and Stefanowitsch's collostructional analysis (Stefanowitsch and Gries 2003, Gries and Stefanowitsch 2004a, Stefanowitsch and Gries 2005, Gries et al. 2005) is another attempt at conciliating corpus linguistics and construction grammars perfectly compatible with that of Lemmens (1998), by qualifying the attraction between lexical items and constructions with a statistical coefficient. However its efficiency is also limited by a tedious manual processing. We will turn back to this method in 1.4.2, by giving the theoretical assumptions underlying this method as well as the computational details.

In a more recent study, Lemmens (2007) attempts at evaluating the degree to which English verbs enter in a constructional alternation in English (here the causative/inchoative alternation), by combining collostructional analysis with a quantification of argument overlap in the alternating constructions. The pilot study should be extended to a larger set of verbs, but doing this manually would not be really feasible. This study, as others, strengthens the idea that research on argument structure constructions crucially needs tools to explore and make use of current corpora.

1.3.4 Towards automatic ASC acquisition

Our work is a clear attempt at reconciliating construction grammar and corpus linguistics, a program that can be called *cognitive corpus linguistics* (hence the title of this section). The ultimate goal of the research assignment underlying this study is to be able to collect constructional information from corpus sentences. More precisely, it aims at making steps towards an effective technique to identify which kind of argument structure construction is instantiated in each clause, and with which arguments.

The result expected after such a program can be described as a constructionally-annotated corpus, where each sentence is described with a constructional representation, that is to say an array of positions occupied by the arguments used in this instance of the construction. Such a rich knowledge of the constructional structure of corpus sentences can prove to be very valuable for further research on constructions. It could be used for various cases in linguistic research where statistical information matters, for example to provide arguments in favour of a usage-based approach to language and to strengthen linguistic theory with the objectivity of a corpus-based approach.

One important thing to note is that no theoretical bias concerning the construction types is being made; we just want to know which syntactic regularities can be derived from scrutinizing a given corpus. Due to the nature of the data that we start from, we cannot make any semantic distinction between formally equivalent cases. As a matter of fact, the puzzle of distinguishing homonymic constructional patterns will be one of our major issues.

1.4 Methodological orientations

1.4.1 Corpus and negative evidence

In 1.3.1, we introduced the methodological approach to language study known as corpus linguistics. As pointed out, the common view nowadays is that corpora and introspection are complementary approaches. While most linguists clearly acknowledge the benefit of using corpora in linguistic analyses, introspective judgements still form the basis of linguistic reasoning, and thus are considered as indispensable. This state of affairs is due to a simple reason: the (generative) test for the explanatory accuracy of one's model is often done through ungrammatical sentences, which obviously are not attested in corpora. One could argue that quantitative data could provide an index for grammaticality, but such an index is actually misleading. Indeed, if some structure is absent from a corpus, one cannot conclude that this structure is ungrammatical, even if the corpus is very large. Some units are naturally more common than others because of their semantics and their felicity conditions. For example, the verb *say* is usually more frequent in corpora than *whisper*, which we can explain by the lower semantic specificity of the former, which allows it to be used in a wider range of contexts. Non-canonical word order in English such as inversion is much less frequent than the usual word order, because the former usually occur in more restrictive contexts than the latter since it is sensitive to discourse factors (*cf.* Birner 1994). As we said earlier, corpora are the only available source of quantitative information, but it appears that such information cannot be used in replacement of introspective judgments. A second major argument against the use of corpora is the suspected futility of quantitative data, *i.e.* the fact that frequencies are skewed by non-linguistic parameters. We can quote Chomsky's classical example: the sentence *I live in New York* is more likely to be uttered and found in a corpus than the sentence *I live in Dayton, Ohio* simply because there are more people living in the former place than in the latter, and thus more people likely to say this sincerely.

However, recent research in corpus methodology have led to new approaches to quantitative data, aiming at rehabilitating corpora for the study of grammatical phenomena. Chomsky indeed has a point: frequencies of occurence can be heavily biased, but this is only true of raw frequencies; this is a case of what Stefanowitsch (2006) terms the "raw frequency fallacy". According to Stefanowitsch, the apparent inadequacy of quantitative data for grammatical analysis is due to an improper methodological treatment; this claim is the starting point of an entire methodology that aims at deriving grammatical knowledge from quantitative corpus data. Concerning negative evidence (the absence of ungrammatical sentences), Stefanowitsch (2006), Stefanowitsch and Gries (2003), Gries and Stefanowitsch (2004a;b) argue that corpora do in fact provide this kind of evidence. They develop a new methodology meant to overcome the raw frequency fallacy and to rehabilitate the use of quantitative data in grammatical analyses. As we know, grammaticality cannot be derived on the sole basis of its frequency or absence in a corpus. The basic claim is that this rough presence/absence distinction should be mapped onto a four-way distinction that takes significance into account. Considering the presence or absence of some linguistic token in some pattern is not enough; one also has to determine whether it is ***significantly** present* or ***significantly** absent* given its frequency in the rest of the corpus. Broadly speaking, if a token T is absent or rare in some pattern P, this absence/rarity will be a significant one if T is very common in other patterns than P. In this view, accurate grammaticality judgments can be derived from corpora, provided a sufficient statistical significance can be assessed, a condition which depends on the size of the corpus.

From these assumptions, Gries and Stefanowitsch designed the *collostructional analysis*, a new method that aims at giving a corpus-based account of grammatical analysis, described in more detail in the next section.

1.4.2 Collostructional analysis

Gries and Stefanowitsch present a series of procedures to compute correlation coefficients between linguistic phenomena in a corpus. Those methods share the goal of collocational analyses, typical of the two last decades of corpus linguistics, but have been adapted to a construction grammar approach, hence the name *collostructional analysis*. In this section, we are going to present three procedures Gries and Stefanowitsch discuss in their publications, namely *collexeme analysis*, *distinctive collexeme analysis* and *covarying collexeme analysis*.

1.4.2.1 Collexeme analysis

Collexeme analysis is the most basic and simple form of collostructional analysis. All other techniques presented in the subsequent sections are more elaborated versions of this one. It aims at quantifying the strength of the relation of a given lexical item towards some constructional context in a corpus. In a collostructional analysis (Stefanowitsch and Gries 2003), constructions are viewed as a set of lexically filled slots. The technique consists in confronting the frequency of occurence of a chosen lexical item in some slot of a construction to the frequency of both phenomena in other conditions. We first have to build what is called a contingency table reporting frequencies, like the following:

	Contruction C	¬ Construction C	
Lexeme L in slot S	$Freq(C \wedge L)$	$Freq(\neg C \wedge L)$	$Freq(L)$
¬ Lexeme L in slot S	$Freq(C \wedge \neg L)$	$Freq(\neg C \wedge \neg L)$	$Freq(\neg L)$
All lexemes in slot S	$Freq(C)$	$Freq(\neg C)$	All patterns

The figures in the right-hand column and the bottom row are called marginal totals and the figure in the bottom right-hand corner is the grand total.

The goal of the experiment is to check whether there is a significant statistical tendency for the lexeme L to occur in the slot S of the construction C under study. The corresponding null hypothesis would be that there is no correlation between the two, *i.e.* a predictable frequency corresponding to the mean expected frequency given the overall distribution, as given by the following formula:

(5) $F(C\&L)_{expected} = \frac{F(C) \times F(L)}{F(all)}$

$F(all)$ equals to the sum of all frequencies, *i.e.* $Freq(C \wedge L) + Freq(\neg C \wedge L) + Freq(C \wedge \neg L) + Freq(\neg C \wedge \neg L)$. If the actual frequency exceeds the expected one, this can be taken as evidence of an attraction between L and C. In contrast, if it is lower than expected, there is a repulsion. However, the total size of the sample affects the strength of the association; the bigger the sample we use, the less probable its data can be accredited to particular variations. In order to properly quantify this attraction/repulsion, Gries and Stefanowitsch argue for the use of a correlation coefficient to measure to what extent the actual distribution fits the null hypothesis. In most of their experiments, they used the Fisher exact test, which is particularly suited for this task because it is not sensitive to the size of the sample, which is an advantage when dealing with usually low-frequency natural language data[2]. Its only drawback comes from its computational complexity, but although it might be slightly longer than other indices like the chi-square test, that should not be a major problem with the help of modern computers. Gries and Stefanowitsch recommend applying a log-transformation to the p-value (*i.e.* the figure yielded by the Fisher exact test), because its "most interesting values are located in the small range of 0.05 to 0" (Stefanowitsch and Gries 2005) and also for the convenience of linguists who are usually not accustomed with the scientific format based on powers of ten. Besides, the log-transformed p-value offers a practical and readable way to check whether a relation is significant or not: as Stefanowitsch and Gries (2005) stipulate, "log-transformed values with absolute values exceeding 1.30103 are significant at the level of 5% (since $10^{-1.30103} = 0.05$)", which means that there is only a 5% probability that the relation is explained by chance and "values exceeding 2 and 3 are significant at the levels of 1% and 0.1% respectively". In social sciences, the hypothesis underlying such a test is generally rejected if this probability is higher than 5%, so we consider attraction values exceeding 1.30103 as conclusive. However, this figure quantifies the strength of an association, but not its direction (attraction or repulsion); the latter has to be deduced from the comparison of the actual frequency with the expected frequency. If the observed frequency is higher than the expected one, the sign of the resulting value has to be changed to a plus.

1.4.2.2 Distinctive collexeme analysis

Distinctive collexeme analysis make use of the same premisses as collexeme analysis, that is to say a 2×2 contingency table giving the distribution of a lexical item in some pattern. However, it is not meant to oppose a construction to all its functional equivalents, but rather two distinct and precisely identified

[2] See Pedersen (1996) for a discussion of this question and a comparison of four correlation coefficients (Fisher exact test, t-test, Pearson's chi-square and likelihood-ratio chi-square test) for the quantification of signifiant lexical relationships in natural language.

constructions having one common slot. The contingency table thus comes as such:

	Construction C_1	Construction C_2	
Lexeme L in slot S	$Freq(C_1 \wedge L_S)$	$Freq(C_2 \wedge L_S)$	$Freq(L_S)$
¬ Lexeme L in slot S	$Freq(C_1 \wedge \neg L_S)$	$Freq(C_2 \wedge \neg L_S)$	$Freq(\neg L_S)$
All lexemes in slot S	$Freq(C_1)$	$Freq(C_2)$	$Freq(C_1 \vee C_2)$

In this case, the resulting value captures to what extent the lexeme L prefers the construction C_1 rather than C_2 in slot S. The rest of the computational details are just the same as in collexeme analysis; as previously, the Fisher exact test is recommended. Such a technique is very useful to compare functionally equivalent constructions (for example, argument structure constructions) and is particularly suited for the study of alternations. It gives an index of whether a verb prefers one construction or the other; thus it can capture subtle differences in the distribution of two near-equivalent constructions, which potentially reflects semantic differences. For example, Gries and Stefanowitsch used this technique to list the most preferred verbs of the ditransitive versus the caused motion (to-dative) construction, and the results confirmed previous non corpus-based analysis.

1.4.2.3 Covarying collexeme analysis

The third and most complex type of collostructional analysis digs more deeply into the interaction of collexemes in constructions. It aims at quantifying the degree of attraction of two lexemes in two distinct slots of the same construction. The contingency table come as follows:

	Lexeme M in slot S_2	¬ Lexeme M in slot S_2	
Lexeme L in slot S_1	$Freq(L_{S_1} \wedge M_{S_2})$	$Freq(L_{S_1} \wedge \neg M_{S_2})$	$Freq(L_{S_1})$
¬ Lexeme L in slot S_1	$Freq(\neg L_{S_1} \wedge M_{S_2})$	$Freq(\neg L_{S_1} \wedge \neg M_{S_2})$	$Freq(\neg L_{S_1})$
All lexemes in slot S_1	$Freq(M_{S_2})$	$Freq(\neg M_{S_2})$	$Freq(C)$

Such an analysis is very specific, because it requires that the slots of a given constructions are identified. However it is a good way to characterize the distributional behavior of a construction, by identifying the most attracted or the most repelled lexeme pairs. In a sense, covarying collexeme analysis can be considered the most advanced generation of collocational analysis.

Stefanowitsch and Gries (2005) argue however that performing this analysis with a Fisher exact test on a 2×2 contingency table would not be different from performing a distinctive collexeme analysis. This raises a methodological problem, since it would not properly do what it is meant for. Stefanowitsch and Gries (2005) suggest two ways by which it could be corrected, one of them involving a more powerful correlation technique called Configural Frequency Analysis on a 3×3 contingency table. However, we will not deal with this issue here since we will not use covarying collexeme analysis in this study anyway. We invite the interested reader to check out Stefanowitsch and Gries (2005) for details.

1.4.3 The hypothesis of a corpus-based grammar

Collostructional analysis offers a glimpse at how corpus quantitative data can be used to reveal grammatical facts. The statistical methods allow it to tackle the raw frequency fallacy pointed out by Stefanowitsch (2006). We argue in favour of Stefanowitsch's proposal that corpus data should play a far more important role in grammar study than it actually does.

Our main starting hypothesis is that we can derive a grammar from the observation of statistical tendencies in a corpus. We argue that the three indices of collostructional analysis are three representatives of a larger family of statistical indices that can be derived from the corpus, following the same principle of quantifying the correlation between two language variables detected in a corpus. This is a very reasonable hypothesis in a usage-based account to language. What we suggest is to take a corpus as the input for language acquisition; since corpora are composed of naturally-occuring data, this process should yield at least some relevant grammatical facts.

In this study, we will attempt to apply this methodology for the acquisition of argument structure constructions. In the next chapter (Chapter 2), we are going to deal with the choice of the corpus and the underlying theoretical issues. We will then provide a quick summing up of how we extracted data from the corpus and what type of refinement we applied to it. Chapter 3 will deal with the theoretical description of the argument structure construction theory, especially focusing on Goldberg's model, and with the consequences for an approach to ASCs acquisition from a corpus. In Chapter 4, we will describe several statistical indices that could be used to providence evidence for ASCs and we will test them on the corpus.

Chapter 2

Corpus data

2.1 Potential issues with corpus data

2.1.1 Levels of annotations for corpora

There exist a large number of corpus resources for English, most of which feature some annotations to provide for linguistic knowledge. In this section, we will discuss four levels of annotations corresponding to the increasing amount of linguistic information that they entail.

Raw text The most basic level of annotation is of course no annotation at all. What makes such data a corpus is that it often consists of carefully collected texts for a specific purpose, compiled, arranged and structured to facilitate their linguistic exploitation (for example, in a concordancer). Few corpora are of this kind; most have at least the first level of actual annotation: PoS tags.

Part of speech Each word in such a corpus is tagged with its syntactic category, most often called part of speech (PoS). This level of annotation is by far the most common one. Unannotated corpus data can be PoS-tagged through the use of a PoS-tagging software like TreeTagger[3], provided the user is willing to allow an error margin, since such software usually adopts a statistical approach which cannot get flawless results. Another methodological issue with PoS-tagging is that there is no consensus regarding the tagset that should be used to tag a given language like English in an exhaustive and linguistically relevant way. The variety and granularity of available tags strongly depends on the goal of the corpus and the theoretical assumptions of its designer. Some tagsets actually reflect traditional grammar categories, while others are influenced by more recent findings (especially those of the distributionalist school) and current theoretical trends, even though there have been some efforts from corpus designers towards genericity, reusability and cross-compatibility of tagsets. Nevertheless, there is still much heterogeneity in this domain, which potentially raises two issues: any software component dealing with corpus data must be tailored to a given tagset and using the same component with another corpus using a different tagset would require an extensive adaptation work, that is to say a partial rewriting.

Lemma The lemma, the canonical form of a word, is important to make significant generalizations and to look for lexical items regardless of their specific form in context. It is often available when the corpus is PoS-tagged; if it is not, it can easily be obtained with the help of a lemmatizer which makes use of some lexical database (like WordNet). If the PoS is given, then the lemma can be determined in a reliable way, since homonymy between inflected forms is often disambiguated when the part of speech is known. For instance, the word *breaks* can be the third person present simple of the verb *break* or the plural form of the noun *break*; knowing whether it is a verb or a noun thus straightforwardly triggers the disambiguation. However, there do exist a few exceptions (though rare), like in the following famous example:

[3] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

(6) They *saw* a tree.

In this example, the word *saw* can be the present form of the verb *saw*, or the preterit form of *see*. PoS-tagging this word as a verb thus does not remove the ambiguity between two possible lemmas. In this particular case, there is even not enough information in the sentence to disambiguate manually: we would actually need more information from the context. Such cases are so rare that they can easily be solved by a manual correction.

Syntactic structure The next level of annotation consists in providing the parsed syntactic structure of each sentence, usually in the form of trees; hence the name *treebanks* for such corpora. Words in each sentence are arranged in a constituent hierarchy. The design principles of trees can of course be influenced by the theoretical assumptions of the treebank designer. Various design options can be opted for: binary branching (appealing to the conventions of generative grammar), word compounding (as opposed to full decomposition in graphemic units), traces left by transformations, co-indexing of traces and anaphora, etc. Most syntactically annotated corpora also feature grammatical function tags, especially at clause level. The terminology inherited from traditional grammar, especially including well-known concepts such as subject and object, is acknowledged by most (if not all) corpus designers. However, this terminology does not cover the entire range of dependencies that can be found in language, and there is moreover no real consensus as to which set of functions should be used. The range of functions available in the corpus depends on the choice of the designer.

2.1.2 Constituency and dependencies

While phrase structure trees are usually considered the hallmark of generative grammar, the notion of consitutents is not at all incompatible with a constructional approach. Regardless of the theoretical stance one adopts, evidence for constituency provided by well known syntactic tests (coordination, substitution by a pronoun, cleft, dislocation, questions, etc.) still holds. It is true that construction grammars do not presuppose the existence of phrase categories independently of linguistic expressions accomodating them. However, such an account does not discard the necessity for phrase structure constructions, which should in the end be more or less equivalent to the familiar tree-like XPs.

The aim of this mémoire is to retrieve instances of ASCs from a corpus. Let us assume that this data will in the worst case consist in raw, unannotated text. Of course, in most cases, it will be tricky, if not impossible, to extract in a simple and reliable way the full argument phrases. Consider for instance the following sentences:

(7) I saw children from the school.

(8) I saw a man with binoculars.

Are the expressions *children from the school* and *a man with binoculars* to be taken as one or two complements? In each case the interpretation is different. Moreover, the dependencies between constituents of different levels have to be established if we are to get argument structures; of course, we cannot get accurate dependencies if we have wrongly assembled constituent structures, and vice-versa. The notion of dependency is intimately intertwined with that of constituency; in the previous examples, each parsing solution corresponds to different dependencies, *i.e.* towards the verb or towards the direct object. A crucial issue here will be that of ambiguities of attachment. The following examples illustrate another case of ambiguity, this time between two verbs at different syntactic levels:

(9) I found something to eat on my way home.

(10) I've just planned to train at home.

In (9), the prepositional phrase *on my way home* can be attached either to the verb *found* or *eat*; the interpretation of the sentence is of course different in each case. The same comment can be made about 10, where the phrase *at home* can be attached to *planned* or *train*; note that the latter attachment would be the most likely.

Working with argument structure will require us to actually parse the input data. There do exist a number of syntactic parsers for English, based on various grammar formalisms, but such a solution would not be without problems. There are very few (if any) broad coverage grammars, *i.e.* grammars with a large lexicon and exhaustive rules that can reliably parse many sentences with minimal manual correction. In practice, such a solution would most probably require human intervention to disambiguate problematic cases and to overcome the parser's lexicon deficiencies. There are also techniques to retrieve lexical dependencies from a flat corpus (most likely POS-tagged), like those described in Bourigault and Fabre (2000); *cf.* also section 4.4. However, these *ad hoc* techniques are approximations and are still prone to annotation errors.

While certainly relevant, a detailed discussion of parsing issues largely exceeds the purpose and scope of our study. In order to guarantee a reasonable level of quality and accuracy for our results, we limit ourselves to data that has already been parsed, such as a treebank.

2.1.3 Meaning

Another issue that poses problems for the type of corpus-based study that we envisage is that construction grammar, unlike formalist approaches, emphasize the role of meaning. Meaning is one of the two basic facets of constructions, as the following definition given by Goldberg (1995:4) shows:

> C is a construction iff C is a form-meaning pair $< F, S >$ such that some aspect of F or some aspect of S is not strictly predictable from C's component parts or from other previously established constructions.

Obviously, argument structure constructions are constructions in the first place, and thus must comply to this definition involving the meaning of the whole clause. However, we do not have access to the meaning of linguistic expressions in corpora, we only have access to their form. In order to check for argument structure constructions in a corpus, it will be necessary to find some way to overcome this drawback. This issue is very common when dealing with corpora, and many researchers have tried to overcome it by postulating rules of thumb or theoretical principles about the form-meaning correspondences. A popular choice in statistical NLP is the assumption that there is a strong correlation between the meaning of a word and its distribution, *i.e.* to the set of syntactic contexts it appears in, which means that two words with similar distributions are likely to have very close meanings. In most corpus-based studies, it is assumed that for a huge amount of data, a purely form-based account is a good approximation and that specific semantic phenomena are only marginal.

In the constructional approach, the notion of semantic coherence allow, however, to derive two relevant observations:

1. Instances of the same construction will tend to get their arguments from a finite and identifiable semantic class;

2. Arguments of a given instance (including the verb itself) should display some semantic compatibility that allows them to be used in this construction.

Gries and Stefanowitsch clearly showed that arguments of an ASC are linked by complex semantic relations, when they are not co-dependent at all, displaying in a way the behavior of collocations. Their corpus studies (Gries and Stefanowitsch 2004a, Stefanowitsch and Gries 2005) clearly illustrate the principle of semantic coherence and show that corpora can indeed provide evidence for it.

2.2 Collecting data

2.2.1 The Susanne corpus

As we said, since our study deals with grammatical constructions and aims at reliably retrieving constituents and dependencies, we need to work with an annotated corpus where such grammatical information is readily available: in other words, a syntactically annotated corpus (a *treebank*).

There exist a number of treebanks of English, most of which are not publicly available. The one million word Penn Treebank is a grammatically annotated version of the Brown corpus and probably the most famous resource of that kind (and also the most expensive). The British English component of the International Corpus of English (ICE-GB; one million words as well) is another example of a good sized treebank, which was automatically annotated with the help of a (quite reliable) LFG parser. However, at the time of the study, we did not have access to these expensive resources; fortunately there is one freely available treebank: the Susanne corpus. This corpus was an initiative by Geoffrey Sampson (Sampson 1995) who sought to develop an innovative annotation scheme that would be comprehensive, explicit and nonpartisan, as clearly stated in the following quotation on Susanne's website[4]:

> The SUSANNE scheme sets out to be:
>
> - comprehensive – covering all features of surface and logical English grammar that are definite enough to be susceptible of formal annotation, and including all phenomena that occur in practice in modern English
> - explicit – if two researchers at separate sites are given the same sample of English and asked to annotate it according to the SUSANNE standards, their annotations should be identical
> - nonpartisan – where aspects of grammar are the subject of theoretical controversy, the SUSANNE scheme aims to embody a neutral analysis which rival theoreticians can interpret in their own preferred terms

The theoretical goals of the Susanne scheme are particularly interesting for us, since we appeal to a model of language with no *a priori* assumptions. Figure 2.2.1[5] gives an example of how the syntactic structure of sentences is annotated in the Susanne corpus. Note that even if the classical tree representation is adopted, like in most theories of syntax since Chomsky, the tree structure has nothing to do with what one can find in generative grammar: there is no binary branching and the constituents are arranged in a natural, part-whole-like way.

The corpus contains approximately 130,000 words (in 6,914 sentences) that were carefully annotated with part of speech, lemma, syntactic structure and grammatical functions. The range of functions have been established by resorting to comprehensive grammar resources, mainly Quirk et al. (1985).

2.2.2 Preliminary processing

We extracted from the Susanne corpus all declarative active clauses, including full sentences, nominal clauses (example (11)), relative clauses (example (12)) and fused relatives (example (13)), which amounts to 6,232 clauses in 4,524 sentences.

(11) One Republican senator told this correspondent *that he was constantly being asked why he didn't attack the Kennedy administration on this score.*

(12) He will be succeeded by Rob Ledford of Gainesville, *who has been an assistant more than three years.*

(13) I'll write *what you tell me to.*

[4] http://www.grsampson.net/RSue.html

[5] Exported from the TIGERSearch treebank querying tool; http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/

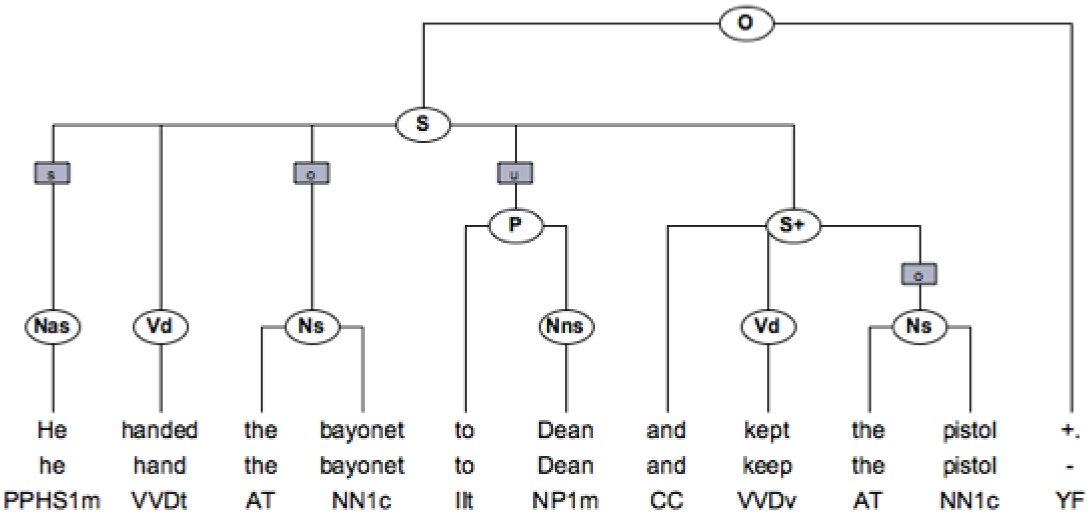

Figure 2.1: Syntactic annotation in Susanne; sentence #4721

Clause-level phrases in each clause could be easily extracted thanks to the syntactic annotations. Phrases first needed to be filtered according to their function; we will return to this in section 3.3.3.

2.2.2.1 Phrases

For each phrase, we extracted the following information:

- The syntactic category (*cat*) ; for convenience and easier reading, we recorded in the main cat field a simplified category information: *N* for nominals, *Adj* for adjective phrases, *P* for prepositional phrases, etc. We left the category as it was given in the corpus in a separate field to keep track of it. We actually did not make much use of this information;
- The grammatical function;
- The governor verb's lemma;
- A pointer to the embedding clause; see 2.2.2.2 below;
- The phrasal head: we construed the phrasal heads as the anchor points of the referent of phrases. This information was gathered in order to get a simplified version of the phrase's content, and to group and compare phrases according to this information. It was required for some indices (see 4.4). To extract heads, we first carefully observed how the structure of phrases was recorded in the Susanne corpus and we designed a simple algorithm that applies the following *ad hoc* rules:
  - For noun phrases, the head is the last noun at the highest level of the constituent structure. If there is none, take the last gerund (which are always tagged as verbs in Susanne, even when they are clearly used as nouns). In case of an anaphoric pronoun, the secondary edge (labelled *CoInd*), if any, is used in order to retrieve the antecedent noun phrase. The basic procedure for NPs is then applied to it.
  - For prepositional phrases, the embedded phrase is first isolated from its preposition. Most if not all PPs in Susanne have two branches at the highest level, the left one containing the preposition (plus some potential modifiers, mainly adverbs), the right one containing the inner phrase. The NP head extraction procedure is applied to this phrase.

  - For determiner phrases (like *a lot of*, etc.), it would be inaccurate to apply the NP extraction procedure, since it would yield the quantifier noun as the head. Hence the NP procedure is applied to the quantified NP, *i.e.* the NP governed by the quantifier's preposition (*e.g. wine* in *a lot of French red wine*, and not *lot* as the rule for standard NP would yield).

- The preposition; in most cases it corresponds to the first final node tagged "preposition" in the left daughter of the PP node. Complex prepositions, such as *out of*, *away from* or *prior to*, were encoded as several prepositions in Susanne; we detected those cases and counted them as a single multiword preposition. In the case of embedded clauses, we recorded the complementizer instead, but we did not actually use this information in the rest of the study.

2.2.2.2 Clauses

For each clause, we recorded:

- The main verb's lemma;

- The full text of the clause, as it was given in the corpus;

- A "short" text of the clause, where all functionally filtered phrases were removed;

- The functional pattern of the clause.

The functional pattern of the clause was recorded in two versions. First, we recorded the surface pattern as it appeared in the corpus, giving the list of functions in the linear order in which they appear and symbolizing the verb by a upper case V. But this encoding is not entirely satisfying because the linear order does not actually matter. It is indeed important to note that ASCs should not specify the linear order of their arguments (at least in English). Goldberg (1995) does not actually say much about how ASCs interact with other clause-level constructions, nor does she deal with word order in the actual sentence. Croft (2001) insists on the fact that an ASC is only a partial description of clausal structure. The main consequence of this assumption regards word order:

> the argument structure of constructions linking participants to syntactic roles will lack specification of the linear order of their elements (p. 197)

Thus the argument structure construction has no influence whatsoever on the resulting word order, which is determined by basic word order conventions (depending on the type of clause: declarative, interrogative, imperative, relative, etc.), as well as by various general principles, like the syntactic weight principle or the information structure ordering. Since constructionally equivalent clauses display different word orders under the influence of various parameters, we should not take the linear order of functions into account when comparing the functional patterns of clauses. For example, while the usual word order for the caused motion construction is Subject - Verb - Object - Oblique, as in (14), it is not uncommon to find the oblique phrase before the direct object when the latter is longer and more syntactically complex than the former (example (15)). The usual word order for the intransitive motion construction is Subject - Verb - Oblique, the oblique can be found sentence-initially in the case of a specific information packaging such as inversion (example (16)):

(14) You take out of circulation many millions of dollars.

(15) Since attack serves to stimulate interest in broadcasts, I added to my opening statement a sentence in which I claimed that German youth seemed to lack the enthusiasm which is a necessary ingredient of anger, and might be classified as uninterested and bored rather than angry. (syntactic weight principle)

(16) To Spahn will go the Sid Mercer Memorial Award as the chapter's player of the year. (inversion, *cf.* Birner 1994)

Besides, we found in the corpus another noteworthy phenomenon. Functions are normally unique in a clause, *i.e.* there can only be one subject, one direct object, etc. However, this does not seem so obvious with adjuncts: there can be several adjuncts of a given type in a clause and this is especially true of time adjuncts, which can often be stacked in complex ways. Since adjuncts phrases have been removed, we should not see this phenomena in our sample; however, as we will see, we have to keep the directional phrases in our sample, even though they are given as adjuncts in the Susanne scheme. Such phrases often occur twice or more in a row when they are used to express complex paths with complementary prepositions, like in examples (17) and (18). Example (19) gives another case, in the form of a directional particle followed by an oblique phrase:

(17) He burst *from* the hot confinement of the room *into* the cold night air.

(18) Other steps would be developed after information drifts *down to* the local level *from* the federal government.

(19) He moved the flights *over against* one wall.

For all those reasons, we decided to record an alternative version of the functional pattern to be able to adequately compare it among clauses. First, word order was ignored: functions where just sorted by alphabetic order (and the verb node was removed), which was also useful for querying since we know exactly what to expect. Second, double directional phrases were reduced to one function tag. The only drawback is that we had to take the preposition and the head lemma from the first oblique phrase, neglecting those of the second phrase.

2.2.2.3 Implementation notes

In the first phases of our investigation, the Susanne corpus was explored with the treebank querying tool TIGERSearch. This software comes with its own query language that can be used to retrieve fine-grained syntactic configurations. Another benefit of this tool is that it can export the query results in the XML format the software is built on (TIGER-XML). The output is fully customizable through the use of XSLT filters. In our case, all relevant clauses were retrieved thanks to the following query:

```
#c:[cat=("S"|"S\@"|"S\?"|"Fn"|"Fn\@"|"Fn\?"|"Fr"|"Fr\@"|"Ff"|"Ff\@")] >
#vp:[cat=/V[^p]*/] & #vp >* #v:[pos=/VV.*/]
```

The query first specifies the clause node (#c), which has to correspond either to a full sentence (S), a nominal clause (Fn), a relative clause (Fr) or a fused relative (Ff); interrogative (?) and appositional (@) variants were included. Passive VPs (p) were left out and the verb node was constrained to lexical verbs only (VV). The results of this query were exported to an XML subcorpus and a XSLT filter was written in order to export all the clause node / verb node ID couples in a raw text file, in order to be able to quickly access the relevant clauses and verbs in their full sentential context.

All the preliminary processing was implemented in Java classes[6]. The TIGER-XML corpus was easily accessed thanks to a Java library, TIGER API 1.8[7], specifically developed for the automatic processing of TIGER format corpora. For each clause node ID in our subcorpus, we retrieved the corresponding syntactic tree via the Java interface. The verb node was used to identify the verb lemma in the VP node. Every clause-level complement was extracted and parsed according to the specifications given in 2.2.2, and all the information was recorded in Java objects, and then automatically saved in a relational database[8] with an object-relational mapping software[9]. A zip file containing the NetBeans project with the full source code has been made available for download at http://florent.perek.free.fr/linguistics/memoire.zip.

[6] We used the NetBeans integrated developement environment; http://www.netbeans.org/.

[7] http://www.tigerapi.org/

[8] MySQL 5.1; http://www.mysql.com/

[9] The Oracle Toplink persistence library, implementing the Java Persistence API specification.

2.3 A potential issue

2.3.1 The size requirement

Before we proceed to our experiments on ASCs in the Susanne corpus, we need to make an important caveat. The corpus-based grammaticality hypothesis we formulated in 1.4 relies on statistical data. As in all statistical studies, any result to obtain will be very sensitive to the size of the corpus, since for any statistical figure holds the rule that the bigger the sample is, the more reliable and comprehensive the results will be. It is particularly true for natural language data. First, it is so likely to be influenced by so many parameters that there always is some risk that such data might be skewed in some way. Secondly, many phenomena in natural language are infrequent and the only way to come across more exemples of those is to increase the size of the corpus. As Gries and Stefanowitsch note, the collostructional analysis as well as the underlying hypothesis rely on the assumption that the corpus is of a sufficient size.

Is the Susanne corpus large enough to draw any significant conclusions about ASCs? 130,000 words in 6,900 sentences is not a very big size for a corpus. In comparison, most other treebanks are at least one million words large, and classical non tree-annotated corpora are quite a lot larger. To defend the relevance of corpora for grammatical analysis, Stefanowitsch (2006) bases his argumentation on a collostructional analysis of the verbal distribution of the ditransitive construction in the one million word ICE-GB corpus. While the results still bring insightful conclusions in favour of Stefanowitsch's point, they show that "even a one-million-word corpus is too small to allow us to identify significant absences for more than a handful of cases" (p. 68), even though he immediatly nuances this by saying that the ditransitive is a "relatively rare pattern". In the face of this result, what to think about the idea that a 130,000 word corpus might provide evidence for argument structure constructions that are likely to be even rarer than the ditransitive? Clearly, definite and conclusive results are unlikely to be the outcome of our small-scale study, and in the dsicussions in chapter 4, we will identify these problems in more detail. Nevertheless, for the purpose at hand, *viz.* the design and implementation of statistical indices, the freely available Susanne corpus proved largely sufficient.

In addition, there is a simple intermediate solution. Instead of trying to get unreliable results for a wide range of underrepresented constructions, we can focus on one or two very common constructions. Those constructions should be easy to spot in the corpus to facilitate the evaluation and should also be representative of the phenomena we try to quantify through the statistical indices. We will then have to extract a subcorpus of all surface patterns that are likely to correspond to those constructions, and see if our indices are appropriate to identify the constructions under study in the subcorpus. From such a method we hope to get three benefits:

1. Even if the corpus is too small for all purposes, we hope to get as much information from it as possible through this method. In the study of a given pattern, it may be unreasonable to draw any conclusions based on statistics from a handful of cases; but if the pattern occurs several hundred times, it becomes possible to compare examples and to assess surface form regularities.

2. On the other hand, although we choose to focus on a limited sample at the expense of other patterns in the whole corpus, we still take into consideration the quantitative data we can get from those patterns. Even if some underrepresented construction C is not quantitatively identifiable, the sentences licensed by C may still be used as a source of information. All the other syntactic contexts in which the verbs from the subcorpus occur should be taken into account.

3. Such a strategy actually fits our purpose, which is the evaluation of statistical indices to quantify ASCs. The limited size of the corpus may then actually turn into a benefit, because it should become possible to test the indices on the whole subsample, which reflects the original balance of the whole corpus.

Before chosing a sample, to which we return in chapter 3, we need to explain how argument structure constructions are supposed to work and illustrate that with a few examples.

Chapter 3

Argument Structure Constructions

3.1 Introduction to Goldberg's model

In 1.2.3, we briefly introduced the concept of argument structure construction and gave the sketch of a definition. Before we can arrive at a full quantitative account of ASCs, we need to define them more accurately than we have done so far. We will attempt to do so in this section, relying on Goldberg's work.

3.1.1 The nature of argument structure

The notion of argument structure construction can be traced back to the earliest works in construction grammar. Goldberg's work can certainly be considered as the most influential in this trend. Goldberg (1995) introduces the concept of argument structure construction as a way to deal with the issues raised by previous approaches and to give a unified account of argument structure in the grammar of a language. Being constructions in the first place, ASCs are form-meaning pairs composed of a semantic pole and a syntactic pole. In Goldberg's framework, the semantic structure of an ASC is composed of a predicate slot and an array of argument roles, which more or less correspond to the classical notion of thematic roles: *agent*, *cause*, *patient*, *recipient*, *instrument*, etc. The syntactic pole specifies how arguments are realized in terms of grammatical functions attributed by the main verb of the clause. Each argument role corresponds to a syntactic function in the syntactic pole: subject, direct object, second object, oblique, etc. The syntactic pole can be formally constrained in order to capture inherent selectional restrictions of the construction. The class of verbs that can occur in the construction is captured by specifying the nature of the relation between the verb and the construction. Most constructions are compatible with verbs whose meaning corresponds to that of the construction or can be construed as an instance of the constructional meaning, but any other type of semantic relation can be specified.

Figure 3.1 displays a graphical representation of the ditransitive construction (adapted from Goldberg 1995:50), exemplified by sentences such as *Mary gave John a cake.* The semantic pole (*Sem*) is in the top of the diagram and the syntactic pole (*Syn*) in the bottom. The construction features three argument roles, an *agent*, a *recipient* and a *theme*, respectively linked to a subject (*Subj*), a first direct object (*Obj1*) and a second direct object (*Obj2*); the linking is symbolized by three arrows. The blank spaces on each arrows symbolize slots that must be filled with the participant roles of the verb (*cf.* section 3.1.3). The dashed line symbolizes a constructional role (*cf.* section 3.1.6). The meaning of the construction, CAUSE-RECEIVE, can be glossed as '*agent* CAUSES *recipient* to RECEIVE *patient*'. The verbs that can occur in this construction are constrained by the possible relations between the verb and the construction, represented by a labelled arrow linking the constructional meaning to the verb slot; here, two possible relations are posited: (i) *means* specifies that the meaning of the verb matches that of the construction, allowing verbs that inherently denote acts of giving, such as *give* or *pass*, and (ii) *instance* specifies that the meaning of the verb can be construed as an instance of that of the construction, allowing specific verbs of transfer such as *send*, *hand* or *ship.*

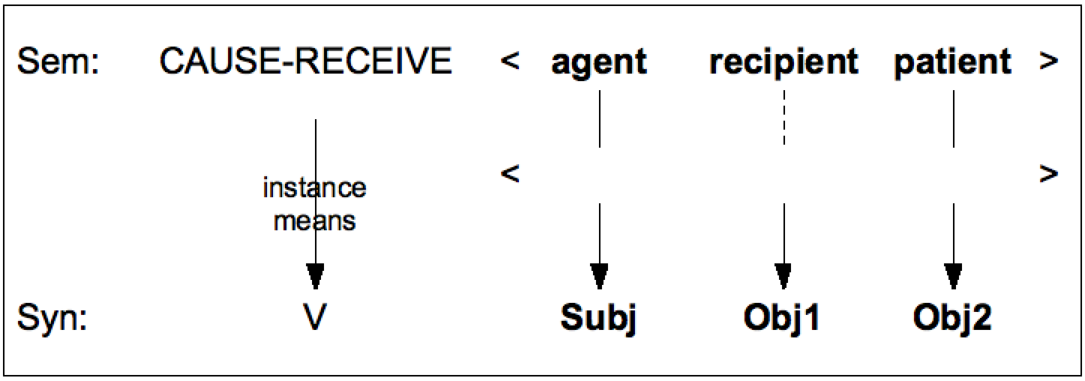

Figure 3.1: The Ditransitive construction

Contrary to traditional lexical semantic accounts that posit several lexical entries (and thus, somehow, several verbal meanings) for each of the argument structures a given verb licenses, the idea of positing argument structure constructions is also an attempt at reducing polysemy by distinguishing lexical meaning (*i.e.* the basic meaning of a verb) from constructional meaning (*i.e.* the actual event encoded by the verb in a given syntactic configuration).

3.1.2 Constructional meaning

All constructionist theories are based on the assumption that constructions, as the basic units of language, are form-meaning pairs. Each construction contributes its own meaning to the sentence. Argument structure constructions are highly schematic, unlike for instance formal idioms (*cf.* Fillmore et al. 1988), like *to beat one's brain out*, whose schematicity is limited to one parameter: an experiencer (*e.g. I've been beating my brains out all day, but I couldn't find the answer!*). In addition to their role as a syntax-semantics interface, they also comply to the following definition of a construction from Goldberg (1995:4) in the sense that they bear an additional semantic value, which might not be provided by other constructions in the sentence:

> C is a construction iff C is a form-meaning pair $\langle F, S\rangle$ such that some aspect of F or some aspect of S is not strictly predictable from C's component parts or from other previously established constructions.

Argument structure constructions are argued to represent what Goldberg (1995) calls “humanly relevant scenes”, *i.e.* they encode basic abstract event types such as something causing something, something undergoing a change of state or someone experiencing something, etc. Such event types are strongly tied to our sensorimotor experience and belong to some finite and seemingly universal set of conceptual categories. They correspond to the central sense of argument structure constructions. In many argument structure constructions, this central meaning can actually be related to that of common “basic purpose verbs”, high-frequency verbs which feature no variation in their meaning as compared to that of the relevant construction. Examples of such a connection for the intransitive motion, the caused motion and the ditransitive constructions are presented in table 3.1 below.

Construction	Constructional Meaning	Basic purpose verb
Intransitive Motion	X MOVES TO Y	*go*
Caused Motion	X CAUSES Y TO MOVE TO Z	*put*
Ditransitive	X CAUSES Y TO RECEIVE Z	*give*

Table 3.1: ASC meaning and basic purpose verbs

These basic purpose verbs are actually argued to be the basis of argument structure generalizations, as evidence from language acquisition shows. Goldberg et al. report that in a language acquisition corpus a small number of those basic purpose verbs in children's speech accounts for a very high number of tokens in the three argument structure patterns analyzed (see Goldberg et al. 2003; 2004, Casenhiser and Goldberg 2005, Goldberg 2005:Part II). They suspect that such unbalanced token frequencies in children's speech could be a clue that those verbs play an important role in argument structure generalization. They thus designed an experiment where children were taught a nonce argument structure through a combination of visual and linguistic stimuli; the experiment was carried out on three groups who were exposed to varying degrees of biased input, *i.e.* the input the first group was exposed to displayed one verb that accounted for a lion's share of the occurences of the pattern, the input of the second featured the same characteristic but far less strikingly and the input of the third group was not biased at all, *i.e.* there was an equivalent distribution for all verbs in the pattern, with no striking difference for one particular verb. The results showed that learners who were exposed to an input with high token frequency of one general exemplar learned the new argument structure more quickly and more easily.

Like many linguistic objects, argument structure constructions are also subject to polysemy. In addition to their basic central sense, they can have extensions that denote more complex event types. For example, the central sense of the ditransitive construction can be glossed as *agent successfully causes recipient to receive patient*. The central ditransitive is thus compatible with (a) verbs of actual giving: *give*, *pass*, *hand*, *serve*, *feed*, etc.; (b) verbs of causation of ballistic motion: *throw*, *toss*, *slap*, *kick*, *poke*, *fling*, *shoot*, etc.; and (c) verbs of "continuous causation in a deictically specified direction" (Goldberg 1995:38), like *bring* and *take*.

(20) The delivery boy handed Susan a packet.

(21) His brother kicked him the ball.

(22) My mother brought us milk and cookies.

But many other verb classes can felicitously license the same pattern:

(23) Mary promised her son a new bike.

(24) The guard denied us entry to the room.

(25) The saleswoman reserved Mary a pair of these new fancy boots.

(26) Her boss allowed Sarah a day off.

(27) Irene knit her grandson a pullover.

(28) Would you grab me a pen, please?

None of the examples (23) to (28) encode an actual transfer of the patient to the recipient. The ditransitive can accommodate verbs of creation (example (27)) and obtaining (example (28)); in both cases the actual meaning is that of an intended transfer, but the sentence does not presuppose that the recipient actually receives the patient. The ditransitive can also denote a transfer that will only occur in some future point in time (example (25)), or if some conditions of satisfaction are met (*i.e.* if Mary keeps her promise, in example (23)). Example (26) exemplifies indirect causation of the transfer. The pattern can even denote a denied transfer (example (24)). In these examples, the basic meaning of a successful, actual transfer is overridden by more complex and subtle semantics. We can gloss the constructional meaning in examples (23) through (28) as follows:

(a) conditions of satisfaction (in the sense of Searle 1969) imply that agent causes recipient to receive patient (example (23))

(b) agent causes recipient not to receive patient (example (24))

(c) agent acts to cause recipient to receive patient at some future point in time (example (25))

(d) agent enables recipient to receive patient (example (26))

(e) agent intends to cause recipient to receive patient (examples (27) and (28))

These are the extensions of the ditransitive basic meaning, as Goldberg posits them (1995:38). One could actually argue that the basic meaning is not overridden but rather modulated and that these modulations are contributed by the verb itself. In his attempt to provide a formal representation in a unification-based grammar, Kay (2005) shows that senses (a) to (c) can actually be subsumed into a broader *modal* ditransitive. In those cases, the modal extension of the construction successfully unifies with the verbal frames, which add some modality information to the constructional meaning. Kay's analysis provides two important insights to the discussion:

1. It shows that the constructional categories that are posited hinge on the level of semantic granularity of the construction and on the options available in the language model.
2. A formal account such as Kay's ultimately requires resorting to lexical semantics for an adequate model to be posited.

All in all, the way lexical semantics combines with constructional meaning is unclear in Goldberg (1995). Also unclear is why these meanings cannot just be accounted for by regular metaphors (as in Lakoff and Johnson 1980) construing pseudo-transfers as actual ones, instead of positing independent constructions.

3.1.3 Participant roles

Cognitive approaches to language adopt the view that meaning is embodied in sensorimotor experience and has to be understood as conceptualization (Langacker 1987, Lakoff 1987, Lakoff and Johnson 1980). In many approaches, verbal meaning is determined by the types of syntactic patterns (*i.e.* argument structures) the verb can appear in, appealing to the assumption that every syntactic difference is motivated by a semantic one, undisputed by cognitivists. Cognitive linguistics however emphasize that meaning is conceptualization: in this view, verbs are concepts of some activity or state. In Fillmore's frame semantics terminology, it can be said that a verb *evokes* a certain frame, that is, some conceptualization of common human experience. Frames can be described as rich semantic content involving generic entities and describing an event, and verbs as lexical instantiations of frames. In the lexicon of any natural language, several verbs can evoke the same frame, but they may differ in the perspective they adopt (Fillmore 1977; 1985). To quote a famous example, *buy* and *sell* evoke the same *transaction* frame, but the former is buyer-centred, while the latter is seller-centered. Thus the following examples can felicitously describe the same situation, but it cannot be said that they are synonymous:

(29) Bill bought a car from Susan.

(30) Susan sold a car to Bill.

Further evidence of this phenomenon is provided by the fact that different verbs characterized against the background of the same frame do not always give the same weight to the same participants, like *pay*, another representative of the transaction frame that emphasize the entity used as a payment (usually an amount of money in our Western society) more than the transacted object or the seller. In example (31) below, either the seller (the payment recipient) or the traded object (in brackets) from the transaction frame can be left unexpressed.

(31) Bill paid (Susan) €5000 (for a car).

The traditional argument roles and subcategorization frames from a lexical semantic account are replaced by participant roles in a frame semantics account: verbs do not project arbitrary argument roles, but they draw their participant roles from the frame they evoke. In addition, verbs can also specify which participant roles are profiled and thus obligatorily expressed. For example, *cut* typically involves a cutter and a cuttee; but we also know that things are usually cut with the help of an instrument. While all three participant roles are plausibly drawn from the semantic frame, only the cutter and cuttee participant roles are lexically profiled and obligatorily expressed respectively in subject and object position. The instrument role can be optionally specified in a prepositional phrase, as the series of examples (32) show:

(32)
a. Irene cut the bread.
b. Irene cut the bread with a knife.
c. * Irene cut with a knife.
d. * The bread cut.

Profiling is a lexical property and is thus fixed for each individual verb. One could argue here however that such lexical profiling is just another way of having lexically determined subcategorization frames, and that the frame semantic account has apparently no advantage over a more traditional subcategorization-based approach. To answer such critics, it should be emphasized that the frame semantic account is an attempt at conciliating encyclopedic knowledge with purely linguistic knowledge, and at presenting a model of conceptualized meaning. It views the lexicon as a way of accessing (*i.e. evoking*) wider domains of rich knowledge; what is relevant to linguistics and has syntactic correlates (*e.g.* in terms of obligatoriness) is the *relation* between lexical words and the frames they evoke.

3.1.4 Linking in construction grammar

As explained in 1.2.1, the goal of any argument structure theory is to account for how participants of an event are linked to their grammatical realizations. In our account, linking the surface argument pattern to the argument structure of the construction is viewed as a categorization operation. In other words, linking requires successful categorization:

1. **the predicate** is categorized as a specific instance of the event evoked by the construction, or as a semantically compatible activity

2. each **participant role** given by the predicate is categorized as an instance of one of the argument roles contributed by the ASC

Participant roles are said to be *fused* with argument roles. The fusion of a given surface pattern with an argument structure construction results in the construal of a semantic event structure.

The two types of categorization mentioned above are covered by the semantic coherence principle:

> Only roles which are semantically compatible can be fused. Two roles *r1* and *r2* are semantically compatible if either *r1* can be construed as an instance of *r2*, or *r2* can be construed as an instance of *r1*. For example, the kicker participant of the *kick* frame may be fused with the agent role of the ditransitive construction because the *kicker* role can be construed as an instance of the *agent* role. Whether a role can be construed as an instance of another role is determined by general categorization principles. (Goldberg 1995:50)

Semantic compatibility can also be provided by meaning transfer phenomena such as regular metaphor and metonymy. As Lakoff and Johnson (1980) demonstrate, those phenomena are omnipresent in all sorts of speech. Example (33) below involves two regular metaphors, that of INFORMATION SOURCE IS AGENT and that of INFORMATION IS PHYSICAL OBJECT, enabling the subject and object participant to be semantically compatible with the ditransitive construction:

(33) On those dirt roads, my GPS always gives me the wrong direction.

3.1.5 Argument role profiling

While verbs specify which of their participant roles are lexically profiled, argument roles are constructionally profiled. Goldberg's definition of argument role profiling is simply based on a distinction between direct and non-direct grammatical relations:

> Every argument role linked to a direct grammatical relation (SUBJ, OBJ, OBJ2) is constructionally profiled. (p. 48)

The direct grammatical relations listed by Goldberg are subject, direct object and the second object of ditransitive. All others are considered non-direct grammatical relations. The link between profiled roles and direct relations works both ways, *i.e. all and only* roles which are are expressed as direct grammatical relations are constructionally profiled.This distinction between profiled and unprofiled argument roles, which parallels that between direct and non-direct grammatical relations, is motivated by the claim that profiled argument roles have a semantically and pragmatically higher status of salience within the scene encoded by the clause.

The consequence of constructional profiling on argument linking is captured by the correspondence principle (Goldberg 1995:50)

> Each participant role that is lexically profiled and expressed must be fused with a profiled argument role of the construction. If a verb has three profiled participant roles, then one of them may be fused with a nonprofiled argument role of the construction.

In other words, the third profiled participant role of the verb may be encoded by a non-direct grammatical relation.

3.1.6 Verbal *vs.* constructional argument roles

In the description of Goldberg's model, we have not yet introduced how examples like the following classical one can be accounted for:

(34) Irene baked John a cake.

The usual reading of this sentence is that John is the intended recipient of the cake. The problem with this example is that John is certainly not a participant role of bake; indeed, the activity of baking does not intrinsically involve any recipient whatsoever. This is further supported by the fact that this complement is fully optional[10].

To handle such cases, argument structure constructions specify which of their argument roles are obligatorily fused with roles of the verb. Roles which are not obligatorily fused with roles of the verb can be contributed by the construction, in cases where the verb cannot provide a participant role that is suitable for this argument role.

In Goldberg (1995), no terminological distinction is provided for this (it was only symbolized by a dashed line in the construction diagrams the author provides throughout the book).Since argument roles obligatorily fused with participant roles in fact always get contributed by the verb, we choose to call them *verbal roles.* Conversely, roles that are not verbal will be called *constructional roles.*

3.2 Constructions under study

We decided to focus on three common constructions in English: the intransitive motion construction, the caused motion construction and the conative construction. In this section, we introduce these three constructions by giving a few semantic and syntactic details. We do not intend to present here a thorough investigation of each construction; during our exploration of the corpus, we systematically referred to the literature when we needed more precise information in order to build the three samples that we used in the evaluation of the satistical indices. This section is primarily meant to present basic syntactic and semantic properties of each construction, so as to make the reader familiar with them.

[10] Of course if it is omitted, the transfer reading does not apply however.

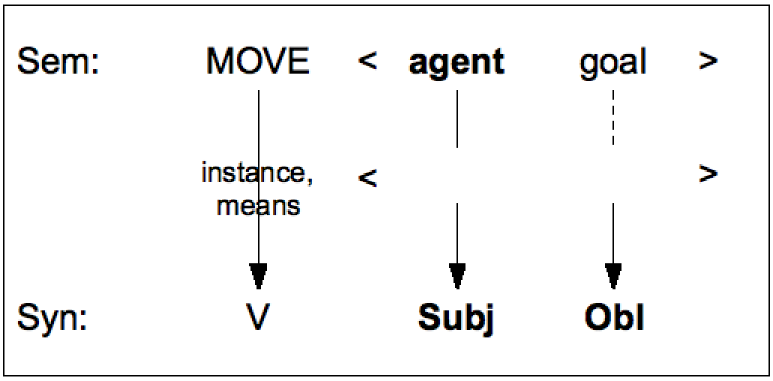

Figure 3.2: The Intransitive Motion construction

3.2.1 The intransitive motion construction

The intransitive motion construction encodes a basic motion event. It links two argument roles, an agent and a goal, respectively to a subject and an oblique phrase. The agent argument refers to the entity in motion and the goal argument refers to the path this entity moves along. The meaning of this construction can be roughly glossed as '*agent* MOVES to *goal*'. The diagram in figure 3.2 sums up these characteristics.

This construction is mentioned but not detailed in Goldberg (1995). We assume that the possible relations between the verb and the construction are *means* and *instances*, restricting it to general verbs of motion (example (35)) or verbs of motion specifying a manner (examples (36)). It can also occur with other kinds of verbs (examples (37)), which can be evidence of an extension of the construction. The extensions of this construction have not been described in the literature. Metaphorical uses of this construction are very common, particularly with the metaphor 'STATE IS LOCATION' (examples (38)).

(35) Boys and men go along the riverbank or to the alcoves in the top arcade.

(36) a. They rode to the Rockfork House, a little farther along the opposite side of the street.

b. We walked down the Rue De L'Arcade, thence along beside the Madeleine and across to a sidewalk cafe opposite that church.

c. Lester's hand fluttered to Cabot's shoulder.

d. The ball floated downstream.

e. The buckskin bolted out of the stall.

f. Nothing appalling or horrible rushed upon these men.

g. Many of them have drifted into the cities and towns and seaports.

(37) a. Some of the oldest, most persistent, and most cohesive forms of social groupings have grown out of religion.

b. The guerrillas were swarming from their bivouac at the west end of the enclosure.

c. His mind flicked through the mental pictures he had from the hours of Aircraft Identification.

(38) a. The positive state came into existence.

b. They all flew into action.

3.2.2 The caused motion construction

The caused motion construction is presented in (Goldberg 1995:chapter 7). It links three argument roles, an agent, a patient and a goal, to a subject, a direct object and an oblique phrase respectively. The

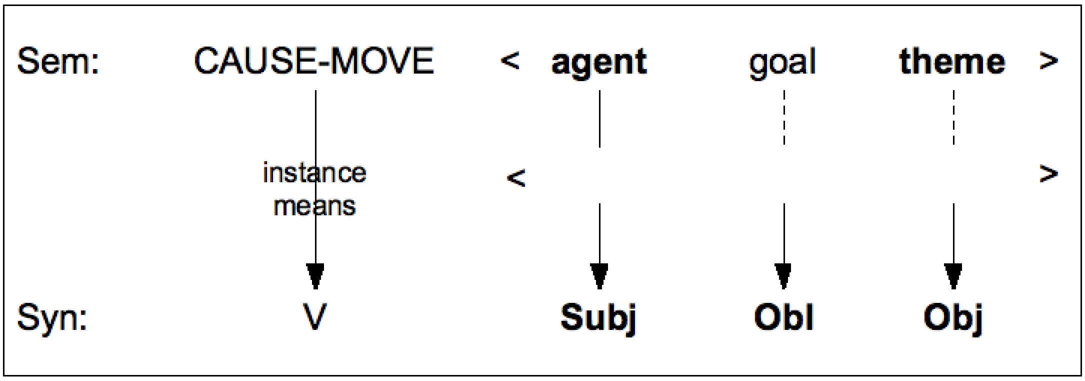

Figure 3.3: The Caused Motion construction

meaning of this construction can be roughly glossed as '*agent* CAUSES *patient* to MOVE *goal*'. The patient and goal argument roles are constructional, they can be contributed by the clause even if there is no corresponding participant role of the verb. The diagram in figure 3.3 sums up these characteristics. According to (Goldberg 1995:chapter 7), the construction has five possible extensions, in other words it has five different meanings that in turn correspond to different verb classes:

1. 'X CAUSES Y to MOVE Z': the central meaning typically occurs with verbs of transfer (*put*, *hand*, *take*, *give*, examples (39)). It can also occur with verbs that do not have either a participant role construable as a theme or a goal (or both). In this case the verb typically specifies the act that causes the motion: *kick*, *sneeze*, *shove*, *push* (examples (40));

2. The conditions of satisfaction associated with the act denoted by the predicate entail 'X CAUSES Y to MOVE Z': verbs of communicative acts (*order*, *ask*, *invite*, *beckon*, *urge*, *send*);

3. 'X ENABLES Y to MOVE Z': the verb encodes the removal of a barrier (*allow*, *let*). As Goldberg further explains, "enablement is understood force-dynamically to involve either the active removal of a barrier or the failure to impose a potential barrier [...] enablement that does not actively involve the removal of a barrier is not acceptable in caused-motion expressions" (Goldberg 1995:161–162);

4. 'X PREVENTS Y from MOVING Comp(Z)': the verb encodes the "imposition of a barrier" (*ibid.*:162) (*lock*, *keep*, *barricade*);

5. 'X HELPS Y to MOVE Z': verbs of assistance (*help*, *assist*, *guide*, *show*, *walk*).

In this study we will conflate all these extensions and consider them as one construction, which is basically the case, since extensions of constructions capture the polysemy of constructional meaning.

(39) a. He handed the guard's rifle to Fiske.
b. I've given willful hurt to no man.
c. The lawyer with whom I studied law steered me off the Socialist track.
d. A proposal to raise dog license fees drew an objection from Councilwoman Virginia Knauer, who formerly raised pedigreed dogs.
e. They dragged him inside the building.
f. The bullet flung Gray Eyes from his horse.

(40) a. The younger boy said the blast knocked him out of bed and against the wall.
b. Morgan jerked his head toward the front door.

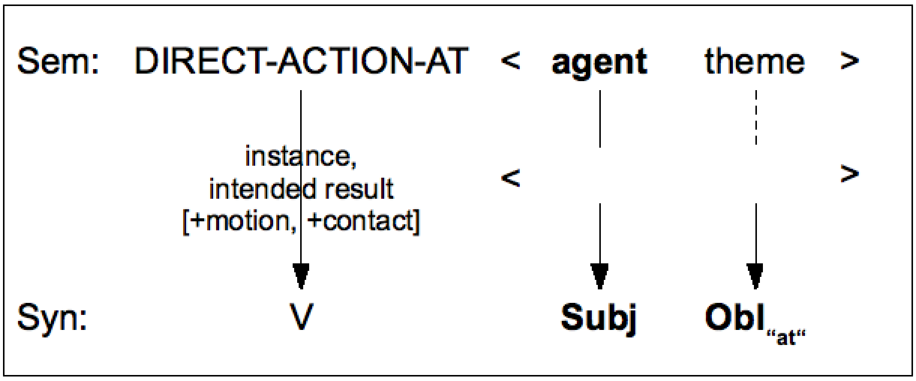

Figure 3.4: The Conative construction

3.2.3 The conative construction

The conative construction is presented in Goldberg (1995:63–64). It links two argument roles, an agent and a theme, to respectively a subject and an oblique phrase with the preposition *at*. The meaning of this construction can be roughly glossed as '*agent* DIRECTS-ACTION-AT *theme*'. The diagram in figure 3.4 sums up these characteristics. The relation between the verb and the constructional meaning can be of two types. In the most typical use, the conative is used to encode an *intended action*, as in the following example:

(41) He scrubbed absent-mindedly at the pans and reflected on how things had turned out.

Verbs occuring in the conative construction with this relation to the constructional meaning can alternate with the transitive construction. The verb denotes the *intended result* of the action. It is restricted to encode both motion and contact.

The second type of relation between the verb and the constructional meaning is more general: it restricts the verb to be construable as an instance of *directed action* (the meaning of the construction). Verbs of seeing are typical in this instantiation of the construction, since verbs like *look* and *stare* can be understood as the projection of one's gaze onto the scene. Various other verbs can also be construed as a directed action.

(42) Brannon looked at Hank Maguire.

(43) And then I became aware that she, too, glanced at me surreptitiously.

(44) Although federal and city narcotic agents sometimes worked together, Sokol continued, rivalries developed when they were aiming at the same criminals.

(45) He had cursed at them and threatened them.

(46) Beyond the stockade rifles began to explode as some of the guerrillas fired at shadows that they imagined were Apaches.

These verbs do not necessarily alternate with the transitive (and usually do not).

3.3 From functional patterns to ASCs

3.3.1 The status of grammatical functions

In 3.1, we saw that the distinction between argument structure constructions in Goldberg's model is based on grammatical functions, such as subject, direct object or oblique. ASCs are apparently not correlated to any particular formal cue. The grammatical functions are the only observable information in the surface form of clauses licensed by ASCs. It is unclear how the constituent structure is parsed and bound to grammatical functions. Apparently, the most straightforward way out would be to assign functions at the lexical level, *i.e.* each verb would assign an array of functions to corresponding syntactic positions of their complements. But such an explanation could not account for arguments that are most plausibly *not* participant roles of the verbs, such as the *recipient* argument in example (34) page 28: if they are not contributed by the verb but by the construction, how could the verb arguably assign them a function? The status of grammatical relations in construction grammar is a challenging question, and we cannot afford to enter into this issue into too much detail. Anyway, grammatical functions seem to be the only way to capture the kind of differences exemplified by (47) to (49):

(47) Susan gave him a kiss.
Subj/NP Verb Obj1/NP Obj2/NP

(48) Susan considers him a fool.
Subj/NP Verb Obj/NP Pred/NP

(49) Susan considers him crazy.
Subj/NP Verb Obj/NP Pred/AdjP

A purely form-based account of argument structure constructions would make no difference between (47) and (48), though they undoubtedly instantiate distinct argument structures. (47) is a typical case of ditransitive construction, whose meaning entails that the recipient argument encoded in the first direct object (*him*) receives the theme argument encoded in the second direct object (*a kiss*). (48) does not entail such a meaning; its argument roles are clearly different: the direct object encodes a theme or stimulus about which the experiencer argument (encoded in the subject) makes some epistemic claim encoded in the second constituent, a predicative. However, their surface forms are fully parallel.

Moreover, a formal account could not predict the generalization we can make over (48) and (49). Both have the same types of arguments and the same semantic relations, but the predicative is encoded by an adjective in the former and by a noun phrase in the latter, which breaks the formal parallelism. Note that the sentence *Susan considers him a crazy person* would be semantically equivalent to (49), and yet formally parallel to (48) (see also Goldberg 2005:21, fn. 2).

This set of evidence brings us to acknowledge that functional patterns are apparently an indispensable cue to the identification of ASCs. Thus, the right interpretation of Goldberg's model for ASC acquisition would be that the set of all possible ASCs corresponds to the set of all observed functional patterns. However, in addition to the controversial status of grammatical functions in Goldberg's approach to argument structure, such a position would pose a number of other problems, which we will show in the rest of this section.

First, even though they have been used throughout the linguistic tradition, there is no consensus as to the range of grammatical functions that are relevant to linguistic expressions. Moreover, we have no guarantee that the set of functions used in our corpus will correspond to functions allegedly relevant for ASCs. So any result obtained under this interpretation of Goldberg's model should be taken with great care.

Second, a functional account would face the notorious problem of what are commonly called adjuncts. They are freely insertable in many (if not all) positions in the clause. Consequently, by their very nature these adjuncts cannot be part of any ASC. We are not sure what function these complements should be given, if they are to be given one at all. If they do have a function, this type of complements should never be considered as argument phrases. A straightforward solution would then simply be to filter out those functions from those corresponding to genuine arguments.

3.3.2 The argument-adjunct distinction revisited

What makes a functionally-marked token an argument of an ASC? In order to answer that question, we need to discuss the different possibilities that Goldberg's account predicts. Recall that in the lexicalist account, one of the corollaries is that clause-level complements can be sorted in two categories, according to whether they are part of the verb subcategorization frame (*i.e.* selected by the verb) or not: this is the traditional argument-adjunct distinction. In the constructional account, this dichotomy no longer holds because the subcategorization frame is replaced by two levels of selection: that of the argument roles of the argument structure construction and that of the (obligatory) participant roles of the verb. This four-way distinction is illustrated by the following table taken from Goldberg (2005:42):

	Role of argument structure construction	Not a role of argument structure construction
Profiled / obligatory participant role of verb	**(a) Argument of verb and construction** *He* devoured *the artichokes.* *She* gave *him* [*a letter*] She put [*the package*] [*on the table*]	**(b) Argument contributed by the verb** She loaded the wagon *with hay.*
Not a profiled / obligatory participant role of verb	**(c) Argument contributed by the construction** He baked *her* a cake. She kicked *him* a ball. She sneezed [*the foam*] [*off the cappuccino*].	**(d) Traditional adjunct** He baked a cake *for her.* She swam *in the summertime.*

Table 3.2: Possible routes to argument status

The simplest case is when there is a perfect match between the arguments of the ASC and the set of expressed participants roles of the verb (cell (a) in table 3.2). This situation is exemplified by (50a).

(50) a. Mary gave Bill a cake.
b. Mary baked Bill a cake.

In (50a), the agent, recipient and theme argument roles of the ditransitive construction are fused with the semantically compatible participant roles contributed by the verb: the giver, the givee and the given object. It is exactly parallel to (50b), but in the latter, the verb *bake* does not contribute the same roles; it only contributes a baker, fused with the agent argument of the ditransitive, and a baked object (the result of the baking) fused with the theme argument since it is a suitable theme (*i.e.* it is semantically compatible with the theme argument role). But it would be unreasonable to consider that the recipient role is an intrinsic participant to the process of baking; baking does not normally involve an intended recipient for the baked object. The recipient argument *Bill* is contributed by the ditransitive construction only. Such examples where arguments are contributed by the construction in addition to those contributed by the verb are quite common, with various kinds of verbs and in many languages other than English, notably French; the English examples (50) can be easily paralleled in that language, as exemplified by examples (51) to (54) below. Note that while (52) is not crashingly bad, there are indeed strong variations in the acceptability of this sentence[11]. It is however judged perfectly and unanimously acceptable when the direct object is turned into a clitic pronoun, as in (54).

(51) Marie a donné un gâteau à Jean.
Mary has given a cake to John.
'Mary gave John a cake'

[11] I personally find that sentences like (52) are perfectly acceptable, even though I must admit that it would be more felicitous (and of course more frequent) in colloquial French than in written, formal speech.

(52) ? Marie a préparé un gâteau à Jean.
Mary has prepared a cake to John.
'Mary baked John a cake'

(53) Marie lui a donné un gâteau.
Mary him has given a cake.
'Mary gave him a cake'

(54) Marie lui a préparé un gâteau.
Mary him has prepared a cake.
'Mary baked him a cake'

Another typical example is given in (55).

(55) John sneezed the foam off the cappucino.

The act of *sneezing* involves only one participant (the "sneezer") and the verb *sneeze* is thus typically used in intransitive syntax. In (55), the additional direct object and object phrase are not of course participants to the act of sneezing, but are arguments to the event encoded by the argument structure construction. The sentence can be paraphrased as 'John caused the foam to move off the cappucino by sneezing'.

Cases such as (50b) and (55) were a notorious problem to the earlier accounts to argument structure, which either suggested a proliferation of verbal polysemy (by adding many 'nonsense' subcategorization frames) or more elaborate *ad hoc* explanations (like the *structure argumentale étendue* – extended argument structure – from Leclère 1978) in order to solve them. Construction grammar offers an elegant and coherent way to deal with these (cell (c) in table 3.2).
The next case (cell (d) in table 3.2), where the complement is contributed neither by the ASC nor by the verb, corresponds to the most prototypical example of what previous accounts call adjuncts. Indeed the central definition of an adjunct is that it is not selected by, and thus does not depend on, any other constituent in the sentence. This case is exemplified by the following examples:

(56) Irene baked a cake *yesterday / for John.*

(57) He broke the window *with a hammer.*

The last case to be considered (cell (b) in table 3.2) corresponds to the situation where an obligatory participant role of the verb does not fit into any conventionalized ASC. Such a possibility might sound strange, because it raises the following question: why would the argument structure construction not include this participant role? The answer pertains to the initial definition of a construction: the existence of a construction in the grammar is motivated if and only if its meaning cannot be obtained by the combination of other constructions in the grammar. This is quite clear in the example Goldberg (2005:42) gives for this category:

(58) She loaded the wagon *with hay.*

The prepositional phrase *with hay* encodes the loaded material, *i.e.* a profiled participant role of the verb. Nevertheless, it is conventionally expressed in a way that usually encodes an adjunct phrase (manner or comitative, according to Quirk et al. 1985). According to Goldberg, a sentence such as (58) features an instance of a simple transitive ASC, plus an adjunct construction encoding an additional participant role, and the overall semantic interpretation is enabled by the casual combination of the two. In the rest of this memoire, we will refer to such phrases as *semi-adjuncts*; the status of such a phrase is indeed somewhere in between the traditional notions of argument and adjunct. Moreover, such a pattern does not apparently generalize to many different verbs with a constant meaning, as for instance the caused-motion construction does, which reinforces the idea that it does not encode an independant meaning.

A closer look at how the surface form of clauses is related to argument structure constructions in Goldberg's account thus reveals that there is no direct mapping between functional patterns and ASCs.

We will then have to find some way to determine whether a phrase is contributed by an ASC or falls in the other categories.

3.3.3 Function filtering

The scheme for functional annotations in Susanne is presented in tables 3.3, 3.4 and 3.5, as provided by the documentation of the corpus.

Tag	Function	Example from the corpus
s	logical subject	In the early eighteenth century *this fantastic city*, then the size of London, started to decline.
o	logical direct object	These needs usually concern *the reduction of guilt and some relief of tension.*
i	indirect object	One Republican senator told *this correspondent* that he was constantly being asked why he didn't attack the Kennedy administration on this score.
u	prepositional object	Mr. Nikolais has made a distinctive contribution *to the arts of costume and decor.* The shock therapies act likewise *on the hypothalamic balance.*
e	predicate complement of subject	Mitchell said the statement should become *a major issue* in the primary and the fall campaign.
j	predicate complement of object	One diplomat described the tenor of Secretary of State Dean Rusk's speeches *as "inconclusive".* We must keep the bloodstream of New Jersey *clean.*
a	agent of passive	This phenomenon has been experimentally investigated in detail *by Maecker.*
S	surface (and not logical) subject	*They* didn't seem to be able to think of any. From my wife's experience and other sources, this seems to be rarely encountered in educated circles.
O	surface (and not logical) direct object	I saw *the clergyman* kneel for a moment by the twitching body of the man he had shot, then run back to his position. The thought made *Pamela* shudder.
G	guest having no grammatical role within its tagma	I am naive, *they say*, to make use of such words

Table 3.3: Function tags in the Susanne corpus; Arguments

As we noticed earlier, there is no direct mapping between argument structure constructions and the function patterns that can be observed in the surface form of utterances. Some phrases are just not indispensable for the semantic contribution of the argument structure to be part of the interpretation. We are faced here with one of the most common issues of all syntactic theories: that of the argument-adjunct distinction.

In this section, we will check to what extent the functional information available in Susanne allows us to decide whether such or such type of phrase has to be generalized in an argument structure construction. What we suggest here is that it is possible to reliably perform a first rough filtering based on function tags, but that this filtering is not sufficient and leaves a number of problematic cases that need more elaborated techniques to be worked out.

First, Goldberg's model gives several grammatical relations that we can straightforwardly map onto the functions in the Susanne corpus: namely subject, direct object and second direct object [12]. According

[12] The Susanne scheme make use of the term "indirect object", which corresponds to the first direct object (when there

Tag	Function	Example from the corpus
p	place	Talleyrand passed his New York law office one night *on the way to a party.*
q	direction	Morgan jerked his head *toward the front door.*
t	time	*At once* he started to glance toward the instrument panel.
h	manner or degree	A similar resolution passed in the Senate *by a vote of 29-5.* *Like Pilate*, they had washed their hands. Few writers have *better* understood their deepest selves.
m	modality	Sheriff Felix Tabb said the ordinary *apparently* made good his promise. *Even* the least alteration will change the quality. All Dallas members voted with Roberts, except Rep. Bill Jones, who was absent.
c	contingency	Movements unfold freely because they are uninhibited by emotional bias or purposive drive. *Since he introduces so much modern music*, I could not resist asking how he felt about it. Then I return to the United States *for engagements at the Hollywood Bowl and in Philadelphia.*
r	respect	He had an uneasy feeling about it. With Maria and me, there's never any problem. And it is expressed, at least to their taste, in a perfect form. Whether historically a fact or not, the legend has a certain symbolic value.
w	comitative	The lawyer with whom I studied law steered me off the Socialist track. The separate layers are joined together by hydrogen bonds.
k	benefactive	His goal was to obtain a National League team for this city.
b	absolute	*Admirably written*, it is a perfect introduction to Swedish history for readers of others countries. The anode holder shown in figure 2 was designed *with two goals in mind.* *Being based on so few events*, these results are of dubious validity.

Table 3.4: Function tags in the Susanne corpus; Adjuncts

Tag	Function	Example from the corpus
n	participle of phrasal verb	No amount of ballyhoo will cover *up* the sordid facts.
x	relative clause having higher clause as antecedent	The Irish accent is, *as one would expect*, combined with slight inflections from the French. This material fluoresces under ultraviolet light *which facilitates its sampling and assessment.*
z	complement of catenative	Rhode Island is going *to examine its Sunday sales law with possible revisions in mind.* The landscape kept *repeating itself.* Then he began *to speak about the tension in art between the mess and form.* A man and a girl happen *to meet*; they look straight at the audience, not at each other.

Table 3.5: Function tags in the Susanne corpus; Others

to Goldberg, these functions are unsurprisingly assigned an elevated status, which allows us to predict that such functions are always contributed by an argument structure construction and thus are always part of that construction. Note that they might well be contributed by the construction alone if they do not encode a profiled participant role of the verb.

(59) John sneezed the foam off the cappucino. (Goldberg 2005)

(60) Slowly and thoughtfully, she slipped the ornament into the pocket of her slacks, moved down the stairs and out of the house.

(61) Gavin slipped his arms around his chest and hugged him fiercely.

(62) Once the door was open, they crowded him inside the dark building.

(63) My father ran him off here six years ago.

Therefore, all phrases tagged *s*, *o* and *i* (*cf.* table 3.3) are argued to be intrinsic parts of an ASC. We will not need to perform any further tests on these to provide evidence for argumenthood.

Table 3.4 gives the list of functions that are assumed to be adjuncts in Susanne's scheme. Susanne's documentation indicates that this adjunct typology was taken from Quirk et al. (1985). Since those names are not self-evident (except maybe for professional grammarians), we checked which functions were likely to correspond to adjuncts in Construction Grammar. Note that since both constructions and verbs select the clausal arguments in a constructional account to argument structure, the traditional criterion of verbal selection could not be applied. We finally validated the typology of the Susanne scheme, with the exception of directional phrases. Such phrases are indeed part of several constructions reported on in the literature, like the caused-motion construction and the intransitive motion construction. After a few checks in the corpus, the following functions were considered not to be part of any construction and thus removed from the surface pattern of clauses: time (t), place (p), manner/degree (h), modality (m), contingency (c), respect (r), comitative (w), benefactive (k), absolute (b). In addition, the "guest" (G; see table 3.3) function from the argument typology and the "relative clause having higher clause as antecedent" (x; see table 3.5) were also removed, since they can be considered as a source of "constructional noise"[13].

The remaining functions were considered as potential arguments for ASCs. The "predicate complement of object" (j) is at least part of the resultative construction (*cf.* Goldberg and Jackendoff 2004, Goldberg 1995:chapter 8); since the "predicate complement of subject" (e) triggers similar complex semantic relations, we have reasons to believe that it should be a potential argument too. The prepositional objects (u) were also kept as candidates to argumenthood, since they display some degree of obligatoriness. They can also be semi-adjuncts.

Surface subject and direct object (upper case S and O) correspond to two possible scenarios. They can be complements moved from an embedded clause to the subject or object position of the main clause, which is clearly a generative analysis of the phenomenon. In examples (64) and (65) below, the position that the complement is assumed to occupy in generative grammar is marked by a trace *t*. The other possibility is that of 'empty' subjects (such as *it* or *there*) in the canonical position of a complement moved in a non-canonical position in the case of some structures like extraposition (*cf.* example (66); the extraposed complement is given in brackets).

are two) in Goldberg's terminology. It is true that this term does not sound quite accurate since English does not have a prepositional marking for such complements. On the other hand, it seems appropriate since it is functionally equivalent to complements that do have a prepositional marking in other languages (for example French), and that it is a remnant of an archaic construction of Middle English, where "the indirect noun phrase was formerly marked with the dative suffix, which explains why no preposition is present today" (*cf.* Bybee and Thompson 1997). Goldberg's terminology seems quite incoherent, because the direct object is called *Obj2* (second direct object) in the ditransitive and *Obj* (direct object) in all other constructions. Anyway this terminology issue is a minor one, especially since there is only one ASC that selects this function (namely the ditransitive). We will continue to use the term *indirect object* in the rest of the study.

[13] Term suggested by M. Lemmens.

(64) Therefore, [*the only unknown structural feature*]$_S$ would appear [*t* to be whether the hydrogen atoms are located symmetrically [1] or asymmetrically [3]].

(65) Except for sophomore center Mike Kelsey and fullback Mike Rice, Meek expects [*the squad*]$_O$ [*t* to be physically sound for Rice].

(66) *IT*$_S$ HAS recently become practical [to use the radio emission of the moon and planets as a new source of information about these bodies and their atmospheres].

The subject and object in (64) and (65) could be considered as arguments of the verb and/or of an ASC. But *IT* in (66) would certainly not be considered as an argument of the verb *become*, and many authors would actually treat extraposition and similar information packaging structures as specific constructions, but not as ASCs. As it is clear that generative grammar and constructions grammars do not treat the phenomena exemplified by (64) to (66) in the same way, conflating those two scenarios in a single grammatical function might give unpredictable results. So we will not take patterns containing these functions into account.

3.3.4 Oblique phrases in Susanne

The prepositional objects (*u*) were also kept as candidates to argumenthood, since they display some level of obligatoriness. However it is not obvious why such cases should be separated from the directional phrases, since they all correspond to oblique phrases in general[14]. Indeed, a short survey into the corpus revealed that keeping those two categories distinct raises two problems:

1. Some known ASCs with a directional phrase are split into the two categories. The oblique phrase is sometimes tagged q, sometimes tagged *u*, with no apparent difference in meaning or other parameters than could explain both annotations. One of the most striking example is that of the conative construction (see Goldberg 1995:pp. 63-34). Consider the following examples; the conative construction is instantiated with the same verb, and the resulting clauses both have the same meaning, but the *at*-phrase is tagged *q* in the first case and *u* in the second:

   (67) He looked at the looming hoods of the supply wagons, struck by a new inspiration.

   (68) Fred Rankin looked at him.

   The explanation of such a discrepancy is unclear; it might be an annotation mistake. In any case we cannot relie on this distinction for those cases.

2. Some metaphorical uses of the directional phrase were tagged *u*. This is particularly true of instances of the caused-motion construction, which were easy to spot in the corpus.

   (69) Polyphosphates gave renewed life to soap products at a time when surfactants were a threat though expensive [...]

   (70) The writers' Gold Tee Award will go to John McAuliffe of Plainfield, N. J., and Palm Beach, Fla., for his sponsorship of charity tournament.

Considering these issues about the Susanne scheme, we decided to collapse all *q*and *u* phrases into a single functional category, that we can appropriately call oblique phrases, in the same fashion as most construction grammar accounts (like Goldberg 1995). This category will be tagged *l*.

3.4 Accounting for formal constraints

3.4.1 Constraints on argument roles

Just as, in generative-flavoured grammars, verbs select their arguments according to their subcategorization frames, ASCs in construction grammar must account for argument selection. It thus follows

[14] One could argue that directional phrases do have specific semantics, but we will shortly see that it is not entirely true given our approach.

that ASCs must include a specification of the constraints that govern the selection of their arguments. In the constructional view, selection of arguments by ASCs is a categorization process. The argument roles are seen as prototype categories, and a complement can be licensed as an argument of an ASC if it is an instance of this category, which in turn is expressed in terms of similarity to the argument prototype.

As cognitively grounded as it may be, the categorization explanation is of no use in the context of a corpus. The second source of argument selection, the grammatical functions, is more relevant to us. As we said previously, it is not clear how functions are assigned to the constituents of the clause. Anyway, functions can be used to detect ASCs in corpora with the appropriate annotations. But is function the only criterion for argument selection? We will show that it is not if we are to account for all the cases that have been identified as ASCs.

The caused-motion construction and the intransitive motion construction display a kind of selection that does not necessarily need to be described as categorization. Both select a subject and an oblique phrase; in addition, the former also selects a direct object. The caused-motion construction encodes an event where the agent subject causes the patient direct object to move along a path encoded in the oblique phrase; the intransitive motion construction encodes the movement of the subject along the same kind of path. Given the semantics of each construction, they both share the constraint that the oblique phrase has to encode a path. One could argue that this constraint may correspond to a formal one, since the 'path' meaning highly depends on a directional preposition like *to*, *into*, *from*, etc. This view only holds to a certain extent. The problem is that these prepositions are highly polysemic and do not always encode path; for example, in (71) and (72) below, *in* and *to* are arguably not used in a spatial meaning.

(71) The City Purchasing Department, the jury said, is lacking *in* experienced clerical personnel as a result of city personnel policies.

(72) I must plead guilty *to* a special sympathy for nomias.

Moreover, the oblique phrase selected by these constructions is not always realized as a prepositional phrase: the case of a directional adverb is very common and we can also find a few noun phrases (even though such examples are marginal and/or very idiomatic):

(73) The ball floated *downstream.*

(74) He moved *ahead* carefully, his left hand in front of him, and came to a wooden partition.

(75) Buster would solve that quarterback problem just as we head *that way.*

(76) I heard subsequently that my Uncle and Aunt had dinner in a nearby restaurant in the French Quarter after which he went *home* to get into his costume to keep the date.

(77) As luck had it, he had not gone *twenty feet in the street* before Pat appeared.

This constraint is thus clearly a semantic restriction. Obviously, such a constraint could not be detected straightforwardly in a corpus, since corpora are essentially formal data. This brings us to the last kind of argument constraints: formal constraints. In particular, we will return to oblique phrases as an example, since they display the simplest construable kind of formal constraints. We deal with these in the following section.

3.4.2 Varying degrees of schematicity

A functional pattern with an oblique phrase, like *Subj-V-Obl*, is only a partial specification of the constraints associated with the corresponding construction. The pairs of examples (78) and (79) have been reported in the literature as distinct constructions; however, they respectively license the functional patterns *Subj-V-Obj-Obl* and *Subj-V-Obl*:

(78) a. Someone has already defined the incident as a notifiable accident. (Gries et al. 2005:637)

b. You take out of circulation many millions of dollars.

(79) a. He scrubbed absent-mindedly at the pans [...].

b. Riverside residents would go to the Seekonk assembly point.

Example (79a) is licensed by the conative construction (Goldberg 1995:63–64). This construction is distinct from the intransitive construction, even though they may be metaphorically related. However they have distinct meanings. Verbs appearing in the conative often alternate with a transitive construction (Levin 1993:41–42), with a slight semantic variation: the conative encodes an *intended* action on the object, but it encodes no change of state. For example, consider the semantic difference between those two examples:

(80) a. Filthy Billy shot the sherif.
b. Filthy Billy shot at the sherif.

Example (80a) entails that the sherif is dead; but while example (80b) does not exclude this possibility, it actually says nothing about what change of state the sherif underwent; he may be dead, wounded or completely unhurt. As we said in 3.2.3, the distributions of the conative and the intransitive motion constructions are very different. The conative does not typically appear with verbs of movement; on the other hand, verbs of visual perception (*look*, *peer*, *aim*) are common with the conative and incompatible with the intransitive motion.

The example of the conative construction clearly shows what we called the varying degrees of schematicity. It means that for a given surface functional pattern, there can be several corresponding argument structure constructions that can be distinguished by different formal cues (among others), *i.e.* differences in their internal schematicity in constructional terms. The question is: should we take these differences in schematicity into account for the detection of ASCs in a corpus? Let us review the consequences if we do not. Goldberg's ASC model has another feature we have not mentioned so far, that is inheritance. Goldberg argues that argument structure constructions may be related by inheritance links: the child constructions inherit the formal and semantic feature of their mother. There can also be multiple inheritance, but we will only consider single inheritance. With such an assumption, not taking the formal constraints into account would amount to considering only a large-purpose mother construction that generalizes over the formal and semantic features of all the inheriting subconstructions and that subsumes the entire functional pattern. Goldberg (1995) calls such cases abstract constructions. We could just assume this mapping of all ASCs to a single category. Beside the fact that this easy solution is not the goal we assigned ourselves in this study, we do not see why such a construction would necessarily exist in the speakers' mental grammar, since it has neither a specific form, nor a specific meaning. Moreover, since we appeal to a usage-based model, how could we assess the existence of some mental entity that is not used as such in actual speech? So we must take formal constraints into account.

Chapter 4

Experiments and evaluation

In this chapter, we will discuss several ways to identify argument structure constructions in a corpus through the use of statistical indices, following the methodological commitment we introduced in 1.4. We will suggest indices that should be capable of quantifying argumenthood and schematicity. A index for argumenthood aims at establishing whether a phrase is an argument, a semi-adjunct or an adjunct. Tests for schematicity, in turn, aim at discriminating the construction pertaining to a given functional pattern according to their formal cues; in the case of oblique phrases, these cues will basically be prepositions.

As the results of our corpus analysis show, no single index can account for either of these two facts. Rather, a combination of indices is what will eventually be needed. Our purpose in this section will be to present the indices we designed and evaluate the results when they are applied to our corpus.

4.1 Introduction

4.1.1 Argumenthood, semi-argumenthood and adjuncthood

The first step to identify ASCs in a corpus consists in determining whether a given phrase is contributed by an ASC or not. Of course, such a distinction has to be made only for non direct grammatical relations, since, as we said, direct grammatical relations are always contributed by an ASC. Goldberg's model discussed in 3.3.2 led us to a four-way distinction, which can actually be reduced to a three-way distinction since we do not need to consider whether the complements encode a participant role of the verb; in other words cases (a) and (c) in table 3.2 are conflated. Clause-level phrases can either be:

1. An argument of an ASC; it may either be a participant role of the verb or an argument contributed by the construction only;
2. A real adjunct in the traditional sense, *i.e.* a free element that is not selected by the construction nor by the verb;
3. A semi-adjunct, *i.e.* a profiled argument of a verb that is not contributed by the construction.

Since we filtered the phrases according to their function, there should not be adjuncts in our phrase database. Of the remaining functions, only the directional phrases could be suspected to be adjuncts, but we do not expect to find many that can be considered as real adjuncts which are not selected by the verb. So we rather seek to perform a two-way distinction between arguments and semi-adjuncts that should reveal regular patterns of complementation.

In section 1.4, we presented the methodological commitments we adhere to and the strategy that we intend to use in order to identify ASCs in the corpus, that relies on the hypothesis that grammar can be derived from the proper analysis of quantitative data from corpora. In the rest of this section, we present a first analysis of the plausible statistical behavior of ASCs and semi-adjuncts, according to what we know about their form and meaning. As we will see, there is not much we can conclude from these intuitions, which will lead to designing more elaborated indices.

Semi-adjuncts encode a participant role of the verb, *i.e.* a participant involved in the process denoted by the verb, but they play no argument role in the argument structure construction that licenses the clause. Since the whole syntactic pattern of the clause does not count as an ASC, this means that it does not encode any generalized "humanly relevant scene" the way an ASC does. We can expect that the encoded scene is less general, and thus less likely to appear with a wide range of verbs. So in a sense, an event with a semi-adjunct is more verb-dependent than an ASC, and the semi-adjunct is semantically determined by the verb.

ASCs encode a verb-independent meaning, which allows them to occur with many verbs as long as they are semantically compatible with the meaning of the construction. They should thus have a high type frequency. Patterns with a semi-adjunct, on the contrary, do not carry an independent meaning. The type frequency of a semi-adjunct pattern should be in general much lower than that of ASCs.

Oblique phrases can either be arguments, adjuncts or semi-adjuncts. A specific formal clue comes into play, the preposition. We should make use of this information, since it is one of the most characteristic formal features of ASCs. For example, the conative construction is formally marked by the preposition *at*, the only available choice for the conative. Note, however, that this preposition is not a strictly distinctive feature for the conative, as example (81) shows.

(81) Other Indians were running at the ponies, shrilling and waving blankets. (= they were running in the direction of the ponies)

The prepositional phrase with *at* can be easily construed as a path and a conative interpretation (of intended causation or directed action) is hard to construe with an intransitive verb like *run*, which invites the conclusion that this is an instance of the intransitive motion construction. However, such cases where *at* marks another construction than the conative are very rare (1 out of 29 occurences of *V + at* in the Susanne corpus).

From a given oblique phrase, we have access to a set of cues: mainly the preposition, the governing verb and the functional pattern of the clause. However, on the basis of these cues, it is very hard to make plausible predictions about the statistical behavior of ASCs *versus* that of semi-adjuncts. Indeed, an argument and a semi-adjunct display the same level of obligatoriness and their relative frequency is just a matter of usage. The only difference between ASCs and patterns with a semi-adjunct is a semantic one: an ASC provides an additional meaning whereas the meaning of a semi-adjunct pattern is fully predictable. But as far as form and frequency are concerned, they should display no significant difference.

4.1.2 The methodological problem of evaluation

One goal of this study is to try to find statistical indices that would be valid for the acquisition of argument structure construction in a corpus. To achieve this goal, we will have to test these indices on our corpus and check whether they conform to the expectation and to what extent. When doing so, we will face a severe methodological problem . As we will elaborate below, this reveals a major methodological problem pertaining to the construction grammar framework, *i.e.* the absence of simple criteria to identify constructions. We will present a possible solution at the end of the section.

The evaluation task in computational linguistics and NLP, and in automatic language learning applications in particular, usually appeals to a generic model. First, a sample of the automatically acquired linguistic resources, such as formal grammars, lexicon, subcategorization frames, semantic information etc., is constituted. For corpus processing applications, this is usually done with a test corpus. Next, this sample is compared to the expected results. These expected results are usually drawn from two sources:

- a *standard*, *i.e.* a pre-existing repository that is considered valid and acknowledged by the linguistic community, *e.g.* a dictionary, a grammar book, the work of a linguist etc.;
- a manual annotation of the test sample by human evaluators.

For example, Schulte Im Walde (2006) designed a system to sort verbs into semantic classes on the basis of the subcategorization frame they occur in; the resulting semantic classes were compared to Levin's (1993) semantic verb classes. Meyers et al. (1996) evaluate the validity of their criteria to distinguish arguments and adjuncts by systematically annotating the arguments and adjuncts in a test corpus and then comparing these to those given by student lexicographers for the same sample. The argument-adjunct disambiguation metrics used in SYNTEX developed by Fabre and Frérot (2002) (*cf.* 4.4.2) were tested via a combination of both methods: first, a randomly selected sample was evaluated by a fellow linguist; then the automatic annotation on the same sample was compared to the subcategorization frames reported in the *Trésor de la Langue Française* dictionary.

Each evaluation method has its own advantages and drawbacks, and should be judged according to the specificities of the phenomenon under study. In the case of argument structure constructions, there is currently no true standard, *i.e.* a generally acknowledged list of ASCs. Of course such a list could be drawn up selecting the constructions already posited in the literature, but there is no guarantee that it would be exhaustive. Besides, drawing up such a list is precisely one of the goals underlying the automatic acquisition of ASCs in corpora.

Moreover, resorting to human judgements seems haphazard too, given the lack of intuitive judgements about ASCs. Assessing a grammaticality judgment, a paraphrase relation or syntactic dependencies seems relatively easy and intuitive, at least in most cases. But deciding whether a given pattern is an ASC or not is a far more complex task: indeed, linguists who posit the existence of an ASC always do so at the expense of an elaborated argumentation that relies on formal and semantic grounds. The cognitive reality of such a construct is then often demonstrated on the basis of corpus and/or elicited data. Gries et al. (2005) provide an excellent example of such an argumentation, about the so-called *as*-predicative construction exemplified by (82) and (83).

(82) She regard her clients' business as confidential (Gries et al. 2005:637)

(83) All governments want to treat arms sales as their own prerogative. (*ibid.*:640)

Appealing to Goldberg's definition of a construction, Gries et al. suggest as a criterion that "for an expression to qualify as a construction, [...] it cannot be compositionally derived in both *form* and/or *meaning* from other constructions available in the language" (p. 639). But for such a criterion to be applied, a solid argumentation is needed, which is challenging even in the context of a thorough linguistic study, let alone using it for evaluation purposes. The risk is that the evaluator takes a decision on insufficient grounds. In short, the lack of generally acknowledged objective criteria for ASCs is a major stumbling block for our purposes.

However, even though we cannot identify all the ASCs in the corpus in the current state of research, we are actually able to identify at least a few of them. A number of ASCs have been amply described in the literature, on syntactic as well as semantic grounds (and the latter matters most for us). Besides, recall that our test corpus consists in only a subset of the possible functional patterns, *i.e.* basically *Subj-V-Obl* and *Subj-V-Obj-Obl*. In that subset, we can expect to find at least three ASCs that are very easy to spot: the intransitive motion construction, the caused motion construction and the conative construction, the two former being moreover quite common and thus well represented. All of these elements contribute to allow at least some evaluation.

The Intransitive Motion construction sample, with its 232 items, is the biggest one; the Caused Motion construction sample has 95 items and the Conative construction sample, 27. We should also get a sample of patterns featuring a semi-adjunct, but constituting such a sample raises a problematic issue. We could adopt the same strategy as with constructions, *i.e.* use a well documented example that would be easy to spot in the corpus. One such example, mentioned in Goldberg's work (Goldberg 1995; 2005) is the construction *load+with*, as in *Mary loaded the truck with hay.* We tried to retrieve other examples in the corpus that formally and semantically parallel the schema LOADER *load* CONTAINER *with* THEME; a list of verbs occuring in this pattern can be found in Levin (1993:50–51), *e.g. cover*, *pack*, *spray.* But none of them could be found in the corpus. One possibility would be to search the corpus in search for examples of semi-adjuncts, but we would face the same problem we mentioned before, *i.e.* resorting on the (potentially fallible) evaluator intuition on a distinction that is hard to draw. Furthermore, it addresses the question of under what condition a recurring syntactic pattern becomes an ASC; we will return to this point in our concluding remarks in Chapter 5.

In sum, at this point it is not yet possible to quantify semi-adjuncts, on theoretical grounds (absence of generally acknowledged criteria) as well as practical grounds (insufficient representation in the Susanne corpus). Nevertheless, we will apply all statistical indices to our data (*i.e.* the three construction sample we mentioned before) to see in how far they provide evidence for at least *some* properties of ASCs. In addition, the results will allow us to formulate more elaborate and more accurate methods to automatically extract ASCs from corpora.

4.1.3 A few words about implementation

The calculation of the statistical indices we are about to discuss have been implemented in additional Java classes, in the same conditions as described in 2.2.2.3. The results have been recorded in separate tables in the database. The PAL libray[15] was used for the Fisher exact test.

An important contribution that we added concerns the implementation of collostructional analysis. Since we had to carry out this analysis in a wide range of applications, we decided to implement our own home-made generic classes (based on the PAL library) in order to save coding time. Two components were developed: an interface to access the database and quickly retrieve any kind of frequency and an implementation of collostructional calculation proper. Since we did not need each of the three types of analysis (collexeme analysis, distinctive collexeme analysis and covarying collexeme) for our study, only the first one was actually implemented.

As far as we know, there does not currently exist any Java tool to perform a collostructional analysis[16], we intend to eventually release this library under an Open Source license once the code is finished, so that any computational linguist developing in Java can use it.

4.2 Constructional roles as a cue for ASCs

4.2.1 Attraction to the reduced pattern

Although, as we pointed out, there is seemingly no difference in the statistical behavior of ASCs and semi-adjuncts patterns, there is one property that is unique to ASCs and can be used to provide evidence that a given clause is licensed by an ASC. Some ASCs feature constructional roles, as the oblique argument of intransitive motion and the caused motion constructions, that can be assigned to arguments that do not correspond to profiled participant roles of the verb. In other words, this means that the ASCs themselves can contribute arguments to the clause. A lexicalist account would assume that this implies extending the argument structure of the verb. This means that, considering a given ASC (let us call it C), some of the verbs that occur in C will also occur with other 'reduced' ASCs, *i.e.* ASCs that are composed of a subset of the argument roles of C. Examples (84) and (85) exemplify such cases:

[15] Phylogenetic Analysis Library, a library with many statistical tools initially developed for bioinformatics at the University of Auckland (New Zealand) distributed with the General Public License; http://www.cebl.auckland.ac.nz/pal-project/.

[16] Note that Gries and Stefanowitsch have already developed tools for Perl and R, but they are not publically accessible.

(84) a. The two men *whipped* their horses into town and flung themselves up the steps of the saloon, crying their intelligence.

b. Above me a dark rider was *whipping* his pony with a quirt in an attempt to hurdle the bales.

(85) a. He *burst* from the hot confinement of the room into the cold night air.

b. *Bursting* paper cartridges, he scattered powder beneath the nearest wagon and dumped the contents of the canister upon it.

c. I heard the whir of an axe and a Canadian's face *burst* apart in a bloody spray.

d. The pipes *burst* and they all laughed and stood in ice water to their ankles while they swabbed the bathrooms. (Brown corpus)

The verb *whip* can occur in the caused motion construction as in (84a), which can be paraphrased as *the two men whipped their horses so as to incite them to move into town*; but *whip* is normally a transitive verb, and thus occurs in a transitive pattern as well (84b). *Burst* in example (85), has only one profiled participant role (a burst object; *cf.* examples (85c) and (85d)) but occurs in contructions that contribute other arguments: intransitive motion (85a) and causative (85b). From a purely statistical point of view, they are likely to occur at least as often with the reduced pattern(s) as with the full one.

For a given verb and oblique phrase, a good basis to derive a distinctive cue for ASCs is comparing quantitative data of the full functional pattern and the reduced pattern. If an oblique phrase is a semi-adjunct, the verb should not occur with the reduced pattern (that does not contain the oblique phrase), or at least only incidentally (under some specific pragmatic conditions for example), since semi-adjuncts encode *profiled, i.e. obligatory*, participant roles. The case of ASCs is somehow more problematic. The problem comes from the disjunction between argument roles and profiled participant roles. We can mention again the example of the ditransitive construction with *give* and other verbs of donation *vs.* verbs that do not profile a *recipient* participant role, like verbs of creation (*cf.* 3.1.6). With some verbs, there will be a total isomorphy between the arguments and participants and so those verbs are not likely to occur in the reduced pattern; but in case the argument role encoded by the oblique is constructional (*cf.* 3.1.6), the construction can contribute the argument role on its own, which gives an apparently extended argument structure to the verb. So in the latter case, the verb should occur with the reduced pattern, which arguably shows that the full pattern is probably an ASC. The reduced pattern might be expected to represent the *standard* pattern of the verb; nevertheless, it is not necessarily more likely to occur in the reduced pattern than in the full ASC pattern. However we can expect the occurences of the verb with the reduced pattern to be noticeable, at least significantly more so than if the oblique was a semi-adjunct.

Note that there can well be cases where more than one role is contributed by the construction, for example with *sneeze* occuring in the caused motion construction (*cf.* example (55), p. 34), where both the theme role (direct object) and the path role (oblique) are contributed by the ASC. Since we only consider the possibility of one constructional role (the oblique), the method we suggest here could not account for such cases. However, given that they are quite infrequent and that ASCs usually have no more than three argument roles, the possibility that more than one role would be contributed by the construction alone and not the verb is very small.

To capture whether the oblique phrase is linked to a constructional role, we suggest to calculate for each clause the attraction of the verb toward the reduced pattern. The contingency table for this collostructional analysis is given in 4.1. Pseudo-propositional logic variables and formulae are used to abbreviate the description of the frequencies we have to retrieve from the corpus: F being the functional pattern under consideration, $F - Obl$ is the corresponding reduced pattern, *i.e.* a pattern composed of the same functions as in F but the oblique phrase (Obl). V is of course the verb being tested. Logical conjunction operators are used to combine conditions for frequencies of complex configurations; for example, $(F - Obl) \wedge V$ means 'the functional pattern of the clause is F's reduced functional pattern and V is the verb of the clause'.

	F - Obl	F	
V in Verb slot	$Freq((F - Obl) \wedge V)$	$Freq(F \wedge V)$	$Freq(V)$
¬V in Verb slot	$Freq((F - Obl) \wedge \neg V)$	$Freq(F \wedge \neg V)$	$Freq(\neg V)$
All lexemes in slot S	$Freq(F - Obl)$	$Freq(F)$	All patterns

Table 4.1: Attraction of the verb towards the reduced pattern (*reducedPatternAttraction*)

4.2.2 Evaluation

For each oblique phrase of the intransitive motion and caused motion constructions in the sample, the attraction of the verb to the reduced pattern was calculated. Recall that since the theme argument of the conative construction (expressed by an oblique phrase with *at*) is not a constructional role, such a test is irrelevant for that construction.

The reduced pattern corresponds to an intransitive syntax ([Subj-V] or function list: [s]) for the intransitive motion construction and a transitive syntax ([Subj-V-Obj] or function list: [os]) for the caused motion construction. A strong repulsion (below -1.30103, *i.e.* significant at the level of 5%; *cf.* 1.4.2.1 p. 12) means that the verb is far more likely to occur with all the functions of the construction pattern, *i.e.* there is a total isomorphy between the functions of the pattern and the roles of the verb; no conclusion can really be drawn about the constructional status of the pattern. An attraction coefficient between -1.30103 and 1.30103 means that the verb is as likely to occur with the full pattern as with the reduced one, and an attraction over 1.30103 means that the verb is strongly attracted to the reduced pattern. In that case, the role encoded by the oblique phrase would not be a participant role of the construction, which is an indicator that the clause is licensed by a construction.

This assumption was checked by taking in each sample all the oblique phrases that have a reduced pattern attraction score over -1.30103, which is the pivot value. The results for the intransitive motion construction and the caused motion construction are summed up in tables 4.2 and 4.3 respectively.

For the intransitive motion construction, the indicator accounts for 127 of the 232 tokens (54.74%), and 44 of the 95 tokens (46.31%) for the caused motion construction. These results should be taken with great care, since many verbs can actually be used without one of their obligatory complements when they receive a generic interpretation or when this complement is enough salient in the context (Rice 1988). This is very common in all sorts of speech, so it can potentially bias the results, especially given the size of the corpus. Every verb in the list was checked for its meaning in the *Oxford Advanced Learner's dictionary* (1995 edition). If for any meaning it was not possible to construe any context where an additional participant expressed by an oblique phrase was not presupposed for the state of affairs to be comprehensible, it was considered as a probable case of an omitted complement (and put in italics in the tables). For example, in the case of *enter* and *burst*, the former presupposes a place the subject enters in, while the latter does not need anything else than the subject for a proper interpretation of the situation to be construed. Of course this method is neither formal nor infallible but it is a good start.

Once those verbs have been marked, it can be noticed that the remaining ones conform to our expectations. In the case of the intransitive motion construction, the verbs that are least attracted to the reduced pattern are verbs of motion specifying a manner, *i.e.* verbs than can indeed occur without a directional phrase but are very likely to in common speech, and moreover, verbs that correspond more closely to the central meaning of the construction. In the set of verbs with high attraction to the reduced pattern, non-motion verbs are more common: *single*, *rumble*, *thud*, *flash*, *triple*, *smash*, *double*, *retire*, *play*, *work*. A special mention should made for *break*, which behaves like a motion verb when used with an oblique phrase (*cf.* Lemmens 2006:24) as in example (86), but is a typical inchoative verb in its intransitive use.

(86) The two horses broke from the yard, from the circle of light cast by the lamp still burning in the house, into the darkness.

Attraction for *Subj-V*	Verb(s)	Frequency
1.344	work	2
1.170	play	1
0.898	leave	1
0.678	arrive, disappear	2
0.592	retire, *start*	3
0.539	rise	1
0.279	double	1
0	*penetrate*, smash	2
-0.135	flee, gush, jump, stumble, triple, wander	6
-0.180	withdraw	2
-0.248	walk	5
-0.311	*add*	1
-0.317	bolt, *brush*, burrow, burst, clamber, clatter, dart, dash, *descend*, fasten, filter, flash, flounder, flutter, launch, race, rumble, scuttle, snowball, *splash*, stagger, *stream*, stride, surge, swarm, thud, tramp, tunnel, whip, yank	30
-0.325	hurry, jerk, single	3
-0.345	*return*	6
-0.476	ride	3
-0.488	drop	3
-0.543	*result*, roll	4
-0.628	fall	3
-0.634	crash, creep, *cross*, drift, *enter*, flick, leap, rush, sink, slide, tear	15
-0.951	climb, lean, pour, spread, *stem*, swing	15
-1.030	*break*, fly	4
-1.268	spring, step	5
-1.278	run	9
	Total	127

Table 4.2: Attraction for the reduced pattern in the Intransitive Motion construction

Attraction for *Subj-V-Obj*	Verb(s)	Frequency
11.304	say	3
0.176	carry	1
0	drag, hammer	2
-0.262	push	1
-0.307	grab, kick, shake	3
-0.311	add	1
-0.439	crowd, slap	2
-0.541	concedere, turn, wipe	3
-0.693	lower, slam, whip	3
-0.892	lift	2
-0.971	astound, boost, brush, bounce, *donate*, elevate, free, heave, hook, hurl, inflict, jerk, mount, rip, plant, propel, scatter, *ship*, smash, *transfer*	20
-1.037	run	1
-1.228	back	2
	Total	44

Table 4.3: Attraction for the reduced pattern in the Caused Motion construction

Similar comments can be made about the caused motion construction, though it is not as striking. The bottom of the list is occupied by verbs of transfer (central meaning): *back*, *free*, *hurl*, *donate*, *ship*, *transfer*, etc. The three last verbs are considered suspicious since they should not normally occur without an oblique phrase encoding the 'goal' argument; however, this case clearly shows that even though those verbs should not be in the list at all, they rather appear at the end, which shows that their use in the transitive syntax is more marginal than that of verbs in the top of the list since it is less supported by corpus statistics. Highest on the list are more typically transitive causative verbs: *whip*, *slam*, *turn*, *crowd*, *slap*, *kick*, *shake*, *push*, *hammer*, etc. A special comment should be made about *add*. This verb is normally used with a oblique phrase, since it involves a kind of recipient, *i.e.* something the 'addee' is added to. It is however noteworthy that the attraction of this verb towards the transitive syntax is striking (at least significantly enough). Actually this verb does not behave in the corpus as it would normally do according to the standards of syntax: in the types of speech Susanne features, such as novel and newspaper extracts, where narrations with dialogues and reported interviews or declarations are common, *add* is used very often in a verbal meaning and with a presupposed 'addee' argument in those texts, *i.e.* in the sense of 'add to the discourse' or 'add to the subject matter' (*cf.* (87)), which explains its presence in the list.

(87) "The proposal", Sheets said, "represents part of his program for election reforms necessary to make democracy in New Jersey more than a lip service word". Sheets said that his proposed law would offer state financing aid for the purchase of voting machines, enabling counties to repay the loan over a 10-year period without interest or charge. *Sheets added that he would ask for exclusive use of voting machines in the state by January, 1964.*

In sum, the results we can observe for the intransitive motion and the caused motion construction show that what this index reveals about the selectional preferences of verbs occuring in ASCs conforms to the idea of constructional meaning in Godlberg's model. In the threshold interval we chosed, this index provides evidence that the functional pattern is likely to be an ASC for approximately half the instances of the sample in both cases.

4.3 The role of skewed frequencies in ASC acquisition

4.3.1 The acquisition of ASCs by children

In 3.1.2, we briefly introduced an important finding from language acquisition experiments in a construction grammar framework in support of the idea that the meaning of argument structure construction was strongly correlated to the meaning of one "basic purpose verb", and thus could be approximated to the meaning of that verb. According to Goldberg, such verbs would be the basis of argument structure generalizations; this thesis is one of the centerpieces of Goldberg's (2002) surface generalization hypothesis.

In language acquisition, the surface generalization hypothesis and its body of related claims contradict the well-known chomskyan hypothesis about an innate language ability. The main argument of generativists is what is known as the *poverty of the stimulus* argument, *i.e.* that the input children are exposed to is not enough for language learning (Baker 1979). It is argued that human beings must have some "hard-wired" abilities in order to be able to learn a language. In the study on language acquisition, Goldberg, Casenhiser, and Sethuraman (2003; 2004) tried to counter this argument by providing evidence for elements that could possibly facilitate language learning and overcome the *poverty of the stimulus* issue. In the case of argument structure acquisition, they led a series of experiments to unveil some of the processes at work in language learning.

As we said, they noticed in child language corpora that children's speech features skewed verb frequencies, *i.e.* that one particular general purpose verb[17] accounted for 20 to 40% of each ASC under study (Goldberg 2005:76); for instance, *put* accounted for 38% (99/256) of the occurences of the caused motion construction, while in competition with 42 other verbs. Such a discrepancy might simply be explained by the fact that those verbs were the less semantically restricted in their category. Goldberg et al. added to this a hypothesis about argument structure acquisition: children would be sensitive to the presence of such general purpose verbs that would provide a basis for argument structure generalization, which could make up for the *poverty of the stimulus* problem pointed out by generativists, and be taken as support for the surface generalization hypothesis. They tested this further through several elicitation experiments involving the teaching, through a combination of visual and linguistic stimuli, of a nonce argument structure instantiated by nonce verbs[18] with a specific morphological marking.

As evidenced by striking differences in learning these constructions, the results of the experiments led to the conclusion that skewed input does indeed facilitate the acquisition of ASCs. As they increased the skewedness of input data (*i.e.* by providing in the input a higher frequency of one particular token), the exposed learners acquired a better command of the related argument structure construction, as also shown by the results of the subsequent forced comprehension test (Casenhiser and Goldberg 2005:502–505).

In the light of such evidence, we can predict that such skewed data will be required for the entrenchment of argument structure constructions. We suggest to extend this to automatic language acquisition by the computational system we aim at designing. For a given pattern, it is possible to identify which verb is the most frequently used in the corpus. Of course, the information a computer can gather and process from a corpus is by no mean comparable to the vast and rich experience of a human being exposed to language input. A corpus lacks much information that plays a great role when children learn a language: the context, the visual content of the event, the meaning of already known words, etc. Nevertheless, this study will be a good opportunity to check to what extent we can apply this conclusion about human language learning to machine language learning.

Whether a frequency is biased or not can be determined by a collexeme analysis, by measuring the

[17] Which, as a careful reader can expect, corresponded to the central meaning of the construction, but that is another conclusion.

[18] The use of a nonce verb avoids the possible objection that skewed frequencies are explained by less semantically restricted lexemes.

relative attraction of the most frequent verb towards the pattern. If the input is indeed skewed, this attraction should be very high, which we argue to be a characteristic feature of ASCs. On the contrary, patterns with a semi-adjunct do not have an independent, holistic meaning, which in turn means that they cannot be assimilated to a general purpose verb. The consequence is that the distribution of the pattern should be more or less equally distributed over the different verbs, with no simple token taking a lion's share, or at least less strikingly so than with genuine ASCs. The distribution of patterns with semi-adjuncts should be less skewed and the attraction coefficient closer to 0.

The intransitive motion and caused motion constructions, as we said, require the oblique phrase to express a path, which is a semantic constraint. These constructions are realized by a range of *formal variants*, which means that they can occur with a wide range of prepositions, provided the oblique phrase can be construed as a path. It should be pointed out, however, that Goldberg et al. only took formal cues into account in their corpus survey; they did not refine their results with semantic knowledge as one could expect given the nature of these constructions. In Goldberg et al. (2004), both constructions are assimilated to a formal pattern: a verb followed by a locative complement[19], with an object noun phrase inbetween in the case of the caused motion construction. A prepositional phrase is considered locative if the preposition belongs to a pre-defined list of locative prepositions (*e.g. in*, *toward* but not *with*), which is indeed a formal cue. For our test case, we want to quantify the skewed frequency in the corpus with the intransitive motion, the caused motion and the conative constructions and check if we find the same results as Goldberg et al.'s.

We will use the same strategy to retrieve the most frequent verb for each construction in our samples, *i.e.* the search for the most frequent verb will be based on cues that are directly available in the corpus. A major difference, however, is that functional patterns will be searched instead of purely formal ones, which will allow a more effective and precise retrieval. As Goldberg et al., we assume that constraints on the oblique phrase are captured by lexical restrictions: the preposition in the (most frequent) case of prepositional phrases. But unlike them, we do not want to make an *a priori* selection of a range of prepositions. So, for each clause in our constructional sample, the analysis of the distribution skewedness will be relative to the functional pattern and the preposition. In this definition, $MFV(F, P)$ means 'the most frequent verb occuring in the functional pattern F whose oblique phrase features the preposition P'. For example, if we are to analyze the skewedness of the clause *John stepped into the room*, the analysis will be relative to the functional pattern F = [Subj-Obl] and the preposition P = *into* and we will first have to retrieve the most frequent verb in the distribution of the pattern [Subj-Obl$_{into}$]. For such an approach to be accurate, three hypotheses about the relation between a construction and its surface realizations are presupposed:

1. All formal variants of an ASC have the same MFV. The main point of this hypothesis is that the MFV would be a cue to group all formal variants under the same constructional meaning.
2. For a given functional pattern and preposition couple, there should be only one underlying ASC, there cannot be two conflicting ASCs.
3. For a given functional pattern and preposition couple, there can also be a pattern with a semi-adjunct, but since, as we argued, such patterns do not have skewed frequency as they do not have constructional meaning, these patterns should not have an influence in the retrieval of the MFV and the analysis of the skewedness.

To quantify the skewedness of a distribution, we will calculate the attraction of the MFV of a given syntactic configuration (*i.e.* pattern-preposition) couple to this syntactic configuration. The contingency table of the collexeme analysis is given in 4.4. As usual, propositional logic formulae are used to abreviate the expression of the frequencies to be retrieved. $F \wedge P$ means 'the functional pattern of the clause is F and the preposition of the oblique phrase in F is P'; conversely $F \wedge \neg P$ means 'the functional pattern of the clause is F and the preposition of the oblique phrase in F is any preposition but P'. $MFV(F, P)$

[19] "a preposition phrase indicating location, a particle indicating location (*e.g.*, *down*, *in*), a locative (*there*, *here*), or some combination"

	$F \wedge P$	$F \wedge \neg P$	
V in Verb slot	$Freq(F \wedge P \wedge V)$	$Freq(F \wedge \neg P \wedge V)$	$Freq(F \wedge V)$
¬V in Verb slot	$Freq(F \wedge P \wedge \neg V)$	$Freq(F \wedge \neg P \wedge \neg V)$	$Freq(F \wedge \neg V)$
All Verbs	$Freq(F \wedge P)$	$Freq(F \wedge \neg P)$	$Freq(F)$

Table 4.4: Attraction of V = $MFV(F, P)$ to functional pattern F and preposition P

means the same here as defined above.

At this point, a question arises: what should we do if there are more than one verb with the highest frequency in the distribution of a given pattern-preposition couple; which one should we take into account? This is a pseudo-problem, however, because its answer does not affect the resulting coefficient: the same figure for $Freq(F \wedge P \wedge V)$ yields the same Fisher exact test result. But the real question actually is: what should we do if the distribution is not skewed, and how do we make the computer aware of that unskewedness? Consider for example the following distributions, in descending order of frequency:

Pattern 1	Pattern 2	Pattern 3	Pattern 4
12	3	9	9
4	2	9	8
2	2	8	8
2	1	2	7
1	1	2	6
1	1	1	2
1	1	1	1

While the distribution of pattern 1 is clearly skewed and that of pattern 2 is not, the situation is not so clear for patterns 3 and 4. In pattern 3, there is no highest frequency lexeme, since two lexemes share it; moreover, there are three frequencies far above the others and thus they are possible candidates for skewed input. In pattern 4, there is a verb that has the highest frequency, but it is followed by a tail of several lexemes that occur far more frequently than the two remaining ones. What matters here is that pattern 1 and pattern 2 are likely to correspond respectively to an ASC and a semi-adjunct; a collexeme analysis as the one described above would precisely yield a high attraction in pattern 1 and a low one in pattern 2. On the other hand, such a coefficient calculated on the third and forth distributions would be useless since they are ambiguous. We need another qualitative index that would tell us whether the distribution is indeed skewed or whether it is an unclear case like 3 and 4.

Such an index would pertain to the domain of the qualitative description of statistical data. For the moment, we use a simple calculation: the representativity of the most frequent verb V in the distribution of pattern P ($R_{V,P}$), *i.e.* the proportion of token frequency it accounts for in the full distribution. $R_{V,P}$ is equal to the quotient of two frequencies, the frequency of the verb V in the pattern P and the overall frequency of the pattern P:

(88) $R_{V,P} = \frac{Freq(V \wedge P)}{Freq(P)}$

The evaluation should reveal whether this measure is appropriate. In our ASC samples, both the correlation coefficient and the representativity should be relatively high[20].

4.3.2 Evaluation

In this section we evaluate to what extent Goldberg et al.'s (2004) findings can be used in the task of identifying ASCs in a general purpose corpus such as Susanne. More precisely, we will check the following set of hypotheses:

[20] Note that the two measures are not of the same nature and do not belong to the same scale: the attraction is defined in $[-\infty, +\infty]$ and $R_{V,P}$ in $[0, 1]$.

P	MFV	$Freq(\text{MFV} \wedge F)$	$Att(\text{MFV}, F, P)$	$R_{\text{MFV},P}$	$Freq(P \wedge C)$
to	go come	16	7.78 7.72	9.82%	63
for	*call*	12	15.32	19.67%	7
through	pass	11	18.42	34.38%	22
from	come	9	6.34	16.07%	31
into	come	7	4.46	13.21%	36
toward	move	5	8.09	41.67%	10
in	*believe*	4	5.49	18.18%	4
upon	*depend*	3	5.57	23.08%	4
over	get	2	1.77	15.38%	4
past	move	2	3.58	50.00%	3
down	run	2	3.03	22.22%	8
out of	get	2	1.71	14.29%	8

Table 4.5: Skewed frequencies experiment in the Intransitive Motion construction (C)

1. If a given verb-preposition couple occuring in a functional pattern F corresponds to an ASC, then there should exist one verb whose token frequency in this pattern-preposition configuration accounts for a large part of the occurences of this configuration (*i.e.* 20 to 40%). Given the assumed semantic relationship of this verb towards the ASC, the verb-ASC attraction should be very high.

2. All formal variants of an ASC have the same MFV. The main point of this hypothesis is that the MFV would be a cue to group all formal variants under the same constructional meaning.

3. For a given functional pattern and preposition couple, there should be only one underlying ASC, there cannot be two conflicting ASCs.

4. For a given functional pattern and preposition couple, there can also be patterns with a semi-adjunct, but since, as we argued, such patterns do not have skewed frequency as they do not have constructional meaning, these patterns should not have an influence in the retrieval of the MFV and the analysis of the skewedness.

For each oblique phrase in our three construction samples, we extracted the MFV and its frequency in the clauses having the same functional pattern and the same preposition. For example, for the clause *He handed the guard's rifle to Fiske*, we looked for the most frequent verb occuring in clauses with the functional pattern [Subj-V-Obj-Obl] (function list: [*los*]) and with the preposition *to*, *i.e. give.* Then we computed two figures: $Att(MFV, F, P)$, which measures the attraction of the MFV to the functional pattern and the preposition in question (*cf.* table 4.4), and the representativity of the verb $R_{V,P}$, *i.e.* the proportion of token frequency it accounts for in the full distribution (as defined before). In the preceding example, $Att(MFV, F, P)$ is the attraction of *give* to the pattern [Subj-V-Obj-Obl$_{to}$], and $R_{V,P}$ is the quotient of the frequency of *give* with the functional pattern [Subj-V-Obj-Obl$_{to}$], by the total frequency of that pattern.

For each sample, we isolated the prepositions with a MFV frequency over 1 and retrieved the MFV and its three scores (frequency, attraction, representativity); in addition, we complemented the results by manually checking the frequency of the preposition in the construction under study. The results are reported in tables 4.5 and 4.6. Of course, such a table would have a single row for the conative since its syntax is restricted to the preposition *at*, so we report the results here separately: the most frequent verb was *look*, with 13 occurences on 28, which gave a MFV representativity score of 41.63% and a very high attraction score of 19; obviously, the same preposition appeared in all of the cases.

For the top-ranking preposition (*to*), there are two verbs with the same frequency; *come* and *go*, it should be pointed out that those are two verbs with very close semantics, *i.e.* both verbs of movement

P	MFV	$Freq(\text{MFV} \wedge F)$	$Att(\text{MFV}, F, P)$	$R_{\text{MFV},P}$	$Freq(P \wedge C)$
to	give	9	5.88	9.38%	34
from	*receive*	7	7.62	11.86%	11
into	back	2	3.52	7.14%	13
out of	take	2	1.94	18.18%	5
on	put	2	3.75	25%	4

Table 4.6: Skewed frequencies experiment in the Caused Motion construction (C)

without specification of manner but just a deictic difference. Such a discrepancy with Goldberg's results could be explained by the fact that the latter have been collected on child language corpora, it could then be argued that the vocabulary of adult language is more diversified. The other problem in our results is more serious: in three cases (in italics), the most frequent verb is incompatible with the intransitive motion construction: *call+for*, *believe+in* and *depend+upon* do not encode a motion. Of course it cannot be claimed that the over-representativity of those verbs for those prepositions allow speakers to group these instances with the intransitive motion (hereafter IM) construction: this fact might be the beginning of an argument against Goldberg's hypothesis, though it should be taken with care given the limited size of this corpus and the language type. It could be argued that those two prepositions are more polysemous than the others or just less likely to encode a path; but while this argument would be satisfying with *for*, it would not hold for *in* and *upon*; besides, other prepositions like *to* and *through* are at least as polysemous, especially the former. We do nonetheless have a quantitative criterion pointing out that those cases are indeed problematic: they are the only cases where the MFV frequency is nearly equal or even higher than the overall frequency of the preposition in the construction, which clearly shows that it is impossible that most instances of the preposition-functions pattern would be licensed by the IM construction, since in all other 'unproblematic' cases, the MFV frequency is at least half lower than the preposition frequency in the construction. All in all, it is clear that we cannot relie on formal grounds only to decide whether an instance should be attached to the IM construction: semantic criteria are clearly needed as well.

The twelve prepositions account for 200 instances of the construction (86.2%); the portion drops to 185 (79.74%) if we remove the problematic cases. If we leave these out of consideration, the results confirm our expectations quite well. First, all MFV are verbs of movement and most (except *run*) are possible candidates to general purpose verbs since they do not specify a manner of motion. All MFVs display a satisfactory to very strong attraction, ranging from 1.71 to 18.42. The average value is 5.71[21]. The figures for the MFV representativity are quite satisfactory though less striking: all of them are above 10%[22] and four out of nine are above 20%.

The results for the caused motion construction are far less interesting, a fact one should not be surprised at since the sample is more than twice as small as that of the IM construction (95 *vs.* 232). Even if the data in table 4.6 account for 67 of the 95 instances of the construction, 15 of the 20 prepositions occur in this pattern with a wide range of verbs, none of which occur more than once. In other words, there was no MFV whatever the frequency of the preposition in the pattern. This is of course problematic and can be explained by the size of the corpus and especially by the size of the construction sample. In sum, our small sample does not allow us to confirm that all formal variants of the construction have similar distributions and in particular that their distributions feature skewed frequencies. It would be interesting to test this on a larger sample.

Not surprisingly, the overall most frequent MFV occurs with the preposition *to*, but it is not *put* as in Goldberg's writings but *give*, which has a quite specific meaning compared to the former, which arguably corresponds more closely to the central meaning of the construction. The second MFV in our

[21] The average value of *go* and *come* was taken in the case of *to*.

[22] In the case of *to*, the two verbs should be conflated and their representativity shoudl be summed.

data, *receive*, does not occur in the caused motion construction, as in the cases previously observed in the intransitive motion; indeed a sentence like *I received a package from my parents* does not correspond to the meaning of the caused motion: first, the subject is not the cause of the receiving, but rather a sort of nonvolitional undergoer, and secondly the sentence does not emphasize that the object just moves along a path described by the oblique phrase (here, *from my parents* indicating the source of the package can be construed as a path), but rather that it moves *towards* the subject (even though the path is still involved). Again the inadequacy of the verb is underlined by the ratio between the MFV frequency and the preposition frequency in the construction, which is clearly higher than in the other cases. The occurence of *back* (as a verb) is quite unexpected for an MFV, given its specific meaning; as MFVs should correspond to *general purpose verbs*, they should not have such specific semantic features, such as manner of motion or direction in the case of *back*. *Put* is the least frequent MFV of the CM construction, surprisingly since Goldberg argues it to be its general purpose verb; it occurs only four times (on 95) in the caused motion construction. In general, while the attraction coefficient are quite satisfactory, the representativity ratios are disappointing.

As a conclusion for this section about skewed frequency indices, we turn back to the underlying hypotheses and see how those results relate to them. The good point is that all constructions actually display skewed frequencies; even though some results were not so good, they globally confirm the idea that one particular verb accounts for a large portion of tokens in each formal variant. We can expect that the results would be even more striking with a bigger corpus. On the other hand, a negative point is that the results did not allow to validate the hypothesis that all formal variants have the same MFV. We do find a few formal variants that share the same MFV, but the results globally present a range of dictinct verbs. As expected, MFVs are on the whole verbs with a very general meaning, but we can also find some problematic verbs with quite specific semantics. Again, such inconsistencies with Goldberg's theory of ASCs could be explained by the limited size of the Susanne corpus. And finally, one of the most disturbing issues about those results is that they actually deny the hypothesis that there can be only one underlying ASC per pattern: for a few formal variants, we found skewed frequencies for verbs that were actually incompatible with the construction under study. Whether those verbs are selected by an ASC or are part of another type of surface pattern (as a pattern with a semi-adjunct) has not been determined yet, but what is clear though is that detecting skewed frequencies on the basis of formal and functional cues only cannot reliably provide evidence for the presence of an ASC.

4.4 Measures of verb-oblique dependency

4.4.1 The Syntex analyzer and its new perspectives

In a series of publications about a new syntactic dependency parser, Syntex, Bourigault and Fabre (2000), Fabre and Bourigault (2001), Fabre and Frérot (2002) suggested a new method to quantify the dependency between a verb and a prepositional phrase, which gave interesting results for the argument-adjunct disambiguation. This parser is introduced as a corpus analyzer rather than as a sentence analyzer ("l'analyseur Syntex présente cette différence fondamentale d'être un analyseur de corpus, et non pas de phrases", *ibid.*:135). This means that it makes use of whatever information it can accumulate through the analysis of the whole corpus to support the analysis of individual sentences. Unlike most parsers, Syntex in not based on a formal grammar, *i.e.* a set of predefined rules to determine the syntactic structure of a sentence, even it actually makes use of a subcategorization lexicon[23] in order to establish dependencies between verbs and their arguments. The most interesting specificity of Syntex is that it introduces a new method to resolve ambiguities of attachment of prepositional phrases, *i.e.* the problem of determining whether a given PP is governed by a verb (and if so, which one) or whether it is an adjunct (Bourigault and Frérot 2004). The basic principle is to search the whole corpus first for unambiguous dependencies and compare the results with the potential dependencies in ambiguous contexts in order to disambiguate them. Statistical information about unambiguous contexts is what the Syntex team

[23] Derived from the *Lexique-Grammaire*, a lexicon of French verbs carefully collected in the eighties-nineties by the now dismissed Laboratoire d'Automatique Documentaire et Linguistique (LADL); see Gross (1984).

calls “endogenous resources”, *i.e.* inner resources obtained from no other base than the corpus alone, as opposed to “exogenous resources”, *i.e.* the subcategorization lexicon, that is to say external resources that need to be provided as a complement to the corpus. The strategy of combining these two types of resources significantly improves the resolution of ambiguous attachments, by associating a probability weight to each solution (Fabre and Frérot 2002, Bourigault and Frérot 2004; 2006).

4.4.2 Measuring verb-PP dependency

Fabre and Bourigault (2001) suggest a method to quantify the potential degree of dependency between a verb and a prepositional phrase; this measure is based on quantitative data about the verb, the preposition and the complement (the head of the complement NP). When confronted with an ambiguous PP attachment, the analyzer can provide some hints by looking at the syntactic behaviour of the verb, the preposition and the complement in the rest of the corpus. The productivity of a given verb-preposition pair, taken as the number of distinct complements (in other words, type frequency of the complement), is assumed to be an indicator of a dependency between the verb and the PP.

The method is further detailed and tested in Fabre and Frérot (2002). They further develop the idea that adjuncts display some level of independence towards the verb, since neither their position nor their interpretation is constrained by it, while arguments are constrained by the verb both formally (because of the restrained preposition choice) and semantically (*ibid.*:4; see also Miller 1997, Meyers et al. 1996). For a given verb-preposition-governee triplet, they suggest to combine two measures:

- Governor productivity (*governorProd*): the type frequency of the verb-preposition couple, *i.e.* the number of distinct complements governed by this couple;
- Governee productivity (*governeeProd*): the type frequency of a preposition and complement couple, *i.e.* the number of distinct verbs that govern this couple in the corpus.

Consider for example the following selection of examples:

(89) a. Mary went to the park.
b. John is going to the swimming pool.
c. Bill came to the park with Joe.
d. Sarah went to a park downtown.
e. Joan didn’t go to work today.
f. Sam went to the biggest park in London.

In this sample, the couple (*go*, *to*) occurs four times with three distinct complements: *park*, *swimming pool*, *work*; its governor productivity is thus equal to 3. The sample contains four PPs whose preposition is *to* and whose complement is *park*; the couple (*to*, *park*) is governed by two different verbs: *go* and *come*. Its governee productivity is thus equal to 2.

On the basis of corpus data, Fabre and Frérot argue that triplets with high governor productivity and low governee productivity will tend to be arguments. Conversely, triplets with high governee productivity and low governor productivity will tend to be adjuncts. Fabre and Frérot (2002) test this technique on a geomorphology corpus of French, adopting the following criteria: if the governor productivity is above 2 and the governee productivity is 0, the PP is tagged as an argument; if the governee productivity is above 2 and the governor productivity is 0, the PP is tagged as an adjunct. This procedure yields interesting results (Fabre and Frérot 2002:8–9). The technique was evaluated on two samples of fifty randomly selected triplets, which were compared first to the manual annotations by a fellow linguist, and second to the entries of the TLF[24] dictionary. 88% of the PPs from the argument sample ($governorProd > 2$ and $governeeProd = 0$) were annotated as arguments by the evaluator, and 84% of them were mentionned in the TLF. 72% of the PPs from the adjunct sample ($governeeProd > 2$ and $governorProd = 0$) were

[24] *Trésor de la Langue Française, http://atilf.atilf.fr/tlf.htm*

annotated as adjuncts by the evaluator, and 24% of them were mentionned in the TLF, leaving 76% unmentionned, *i.e.* potentially considered as adjuncts according to the dictionary.

4.4.3 Adequacy for our purpose and discussion

Bourigault, Fabre and Frérot's approach is a classical lexicalist account, in which the subcategorization frame of each individual verbs plays a central role. We can then wonder to what extent their analysis may apply to a constructional approach and how it could help us distinguish arguments from semi-adjuncts.

Our category of semi-adjuncts may pose some classification problems in a two-way distinction between argument and adjunct, since they are somewhere in between the two: while they are obligatorily expressed, they display some features of adjuncthood. First, adverbs can be inserted before them: Goldberg (2002:346) comments that "while placing a clear adjunct before the *with* phrase is not crashingly bad in (90c), it is slightly less felicitous than the corresponding example in (90d)"[25]. Moreover, Goldberg argues that semi-adjuncts behave like adjuncts with respect to the *do so* test (example (91)), in which "arguments are within the scope of *do so* VP anaphora, and adjuncts are outside it" (*ibid.*) and their semantics often closely parallels that of adjuncts.

(90) a. * Pat loaded yesterday the wagon with hay.
b. Pat loaded the wagon with hay yesterday.
c. ? Pat loaded the wagon yesterday with hay. (Goldberg 2002:346)
d. Pat broke the window yesterday with a hammer. (*ibid.*)

(91) a. Liza covered the baby with a blanket and then Henry did so with a quilt. (Goldberg 2002:346), quoted from Rappaport and Levin (1985)
b. * Liza loaded the wagon with hay and then Henry did so with straw.[26]

We are going to test whether this technique to distinguish arguments from adjuncts can also be used to distinguish arguments of an ASC from semi-adjuncts in a constructional approach to argument structure.

Finally, a last caveat: most of the work on Syntex relies on the notion of sublanguage (register) introduced by Harris (1968). According to this idea, the patterns of usage that are common in a given register should be particularly salient in a corpus of that sublanguage. From this assumption it follows that it should be much easier to retrieve characteristic syntactic phenomena from a technical corpus, which is why this technique has first been designed for and tested on specific corpora (though not exclusively). The figures we presented in the previous section come from the evaluation of the technique on a geomorphology subcorpus, which features many patterns of usage that are far more common in the geomorphology register than in standard French. The Susanne corpus, on the other hand, is of a more general register; it is a sample from the Brown corpus and features various kind of texts: newspaper articles, fiction, scientific writings, etc. Thus, the validity of using these indices in our case may be hindered by the non-specificity of our corpus.

4.4.4 Evaluation

For each clause and oblique phrase, viewed as a verb-preposition-complement triplet, we calculated the number of distinct lemmas occuring with the verb-preposition couple (*governorProd*) and the number of distinct verbs occuring with the preposition-complement couple (*governeeProd*). The statistical distribution of these values in the construction samples are reported for the intransitive motion and the caused

[25] Example numbers were adapted to correspond to those given in this document.

[26] Goldberg notices that the apparent unacceptability of this sentence can result from our world knowledge:

> It isn't possible to load a wagon if it is already loaded. Notice (91b) is improved if we assume that the hay Liza loaded is removed before Henry puts straw onto the wagon. (*ibid.*)

However, as she argues further, there is still a difference in acceptability between sentences such as (91a) and (91b).

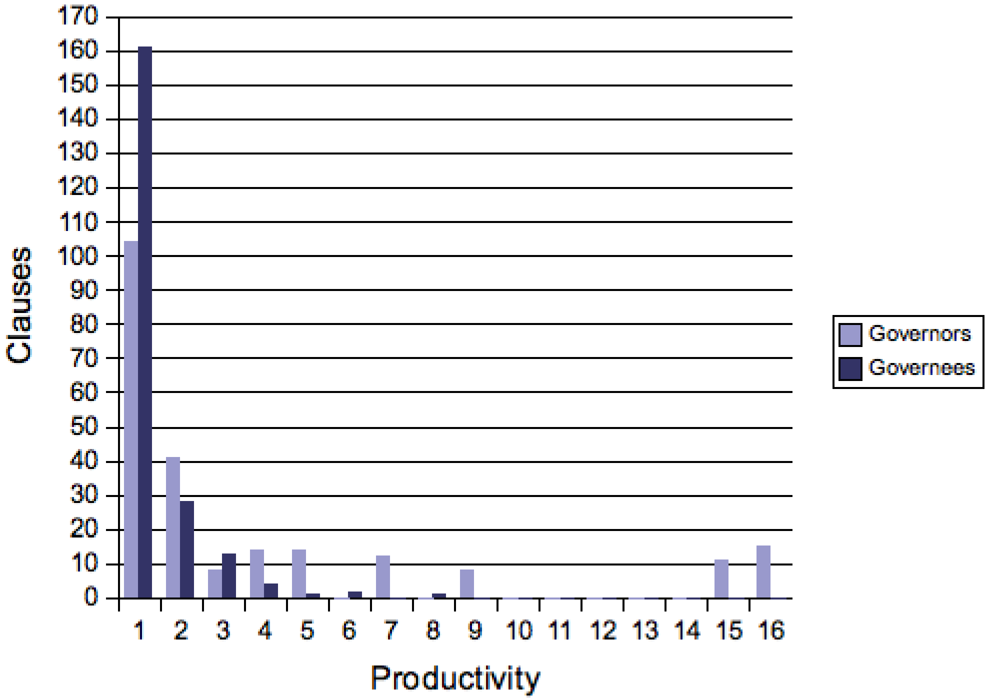

Figure 4.1: Statistical distribution of *governorProd* and *governeeProd* for the Intransitive Motion construction

motion constructions in the histograms in figures 4.1 and 4.2 respectively.

In both cases, there is no striking difference in the distributions of the two indices; unsurprisingly, values are concentrated in the lower ranges (1, 2, 3, 4, 5), which conform to Zipf's law. In the even smaller conative sample, the results were even more disappointing: 14 clauses with a governor and governee productivity of 1, 2 clauses with a governor and governee productivity of 2 (*i.e.* so far the same distribution for governors and governees), and 11 clauses with a governor productivity of 9 (instances of *look*). From this observations it seems that Fabre and Frérot's (2002) hypothesis that high values of *prodGovernor* and low values *prodGovernee* is a distinctive feature of verb arguments does not apply to arguments of ASCs. A strong caveat should be made, however, about the quality of the data, which in turn addresses the problem of the size of the corpus. The productivity of the governees indeed seems to be severely affected by this, at least more strongly than for governors. This is confirmed by the collected data: there are 1194 verb types *vs.* 2877 lemma types[27]. Using a bigger corpus should at least reduce this problem.

4.5 Characterizing schematicity

4.5.1 The distribution hypothesis

In 3.4, we emphasized the need of finding formal constraints of ASCs. One of our goals is to find indices that could be used to discriminate between patterns. In our specific example dealing with oblique phrases, the distinct formal constraints are manifested by distinct prepositions. We mentioned the example of

[27] Prepositions should not be involved in this issue, since they are closed-class words and thus by definition are found in fewer types than open-class like nouns and verbs.

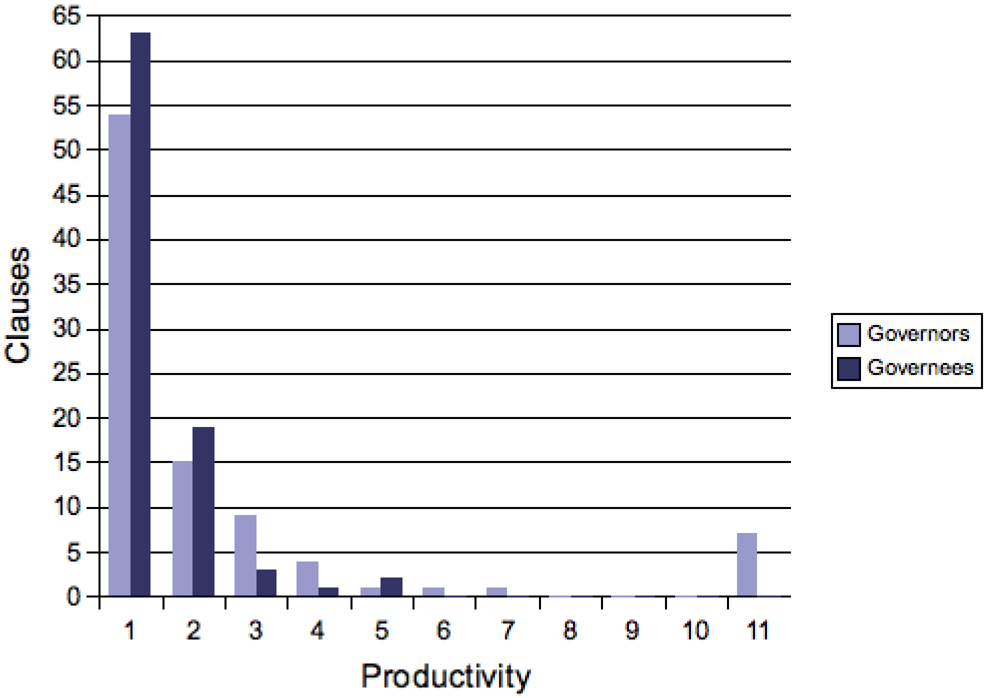

Figure 4.2: Statistical distribution of *governorProd* and *governeeProd* for the Caused Motion construction

two distinct constructions, the intransitive motion construction and the conative construction, which have of course distinct meanings and can be discriminated by distinct formal constraints: in the conative construction, the oblique phrase is constrained by the preposition *at*, while the intransitive motion can occur with a wider range of prepositions expressing a path (*from*, *along*, *beyond*, *to*, *into*, *in*, *onto*, *across*, *through*, *toward*, *over*, etc.).

We suggested in 3.4.2 that the intransitive motion and the conative occur with very different verbs types. Actually, such a discrepancy in the distribution of distinct patterns is totally in line with the idea of constructional meaning developed by Goldberg (1995). Collostructional analysis studies, like Stefanowitsch and Gries (2005), show that the verbs most attracted by a construction are those most closely corresponding to the core meaning of that construction. Gries and Stefanowitsch (2006) provide another kind of evidence with the clustering of verbal lexemes occuring in a given ASC according to their context vector, *i.e.* the set of words the verbal lexeme co-occurs with in the corpus; even though this method approximates the meaning of the verbs, the results clearly show that the most central senses are again those corresponding to the central sense of the construction. Gries (2003) obtained similar results by using a technique to rank instances of the ditransitive and the caused motion constructions according to their prototypicality (in the sense of Rosch 1973) by clustering instances on the basis of a set of formal and semantic cues.

From this it follows that the best approximation of constructional meaning in a corpus is the verbal distribution. Thus, discriminating between two distinct patterns amounts to comparing their distribution. If two patterns with two distinct prepositions have similar verbal distributions, it is unlikely that these patterns correspond to distinct ASCs; conversely, two patterns with very different verbal distributions will be likely to correspond to distinct constructions with distinct meanings. In the case of the discrimination between the intransitive motion and the conative, such a criterion should capture that distinction. We

	$F \wedge P_1$	$F \wedge P_2$	
V in Verb slot	$Freq(F \wedge P_1 \wedge V)$	$Freq(F \wedge P_2 \wedge V)$	$Freq(F \wedge V)$
¬V in Verb slot	$Freq(F \wedge P_1 \wedge \neg V)$	$Freq(F \wedge P_2 \wedge \neg V)$	$Freq(F \wedge \neg V)$
All lexemes in slot S	$Freq(F \wedge P_1)$	$Freq(F \wedge P_2)$	$Freq(F)$

Table 4.7: Attraction of a verb to a preposition P_1 rather than P_2 in a given functional pattern F

nevertheless expect that all prepositions in the intransitive motion should approximately display the same distribution. From this distributional hypothesis we derive quantitative indices that can be used to compare the distribution of a pattern constrained by a preposition P_1 to the distribution of a pattern constrained by another preposition P_2, in other words to quantify to what extent the semantics of the two patterns overlap. To take the example of the conative construction again, comparing the distribution of [Subj-V-Obl] constrained with *at* to that of [Subj-V-Obl] constrained with other prepositions expressing a path is likely to provide evidence that the conative should have an independent status.

Note that the implicit assumption underlying this hypothesis is that there is a direct correspondance between prepositional patterns and constructions. Of course, this may not be true: some prepositions may be used to express a semi-adjunct or may be used in a metaphorical extension licensed by a different construction than the one we expect. We nevertheless assume that for a given pattern, a majority of cases should correspond to only one construct, *i.e.* a single ASC or pattern with a semi-adjunct, while the other possibilities should only be a mere source of noise.

4.5.2 Comparing distributions

4.5.2.1 Collostructional analysis

In the case of quantifying schematicity, collostructional analysis is relevant because we aim at quantifying the difference between the distributions of two patterns. For each verb occuring in the distribution of the functional pattern F constrained by either P_1 or P_2, we suggest to use a form of distinctive collexeme analysis to compute a coefficient capturing the attraction of the verb to P_1 rather than P_2 in the context of the functional pattern. The corresponding contingency table is given in 4.7. As usual, logic formulae are used as a shorthand notation; $F \wedge P_1$ means 'the functional pattern of the clause is F and the preposition of the oblique phrase is P_1'. The resulting coefficient will give an account of whether the verb V is attracted by the pattern with P_1 rather than by the same pattern with P_2. The final index is obtained by computing the average value of the coefficients for all verbs in the distribution of $F \wedge (P_1 \vee P_2)$, *i.e.* for all verb types occuring in F with either P_1 or P_2. Consider for example that we want to compare the distribution of the prepositions P_1 = *to* and P_2 = *from* in the functional pattern [Subj-V-Obl]. In an hypothetical corpus, [Subj-V-Obl$_{to}$] occurs with *come*, *go*, *walk* and *jump*; and [Subj-V-Obl$_{from}$] occurs with *come*, *jump* and *fall*. For each of the five distinct verb types (*come*, *go*, *walk*, *jump* and *fall*), we would have to fill a contingency table by calculating the frequencies as given in table 4.7. The resulting value capturing the similarity between the two distributions will be obtained by calculating the average value of all five attraction coefficients calculated on the contingency tables.

4.5.2.2 Measures of distributional overlap

Lemmens (2007) sketches another way of evaluating the degree to which a given verb enters in a constructional alternation. The study is based on a combination of collostructional analysis with two new measures that are meant to quantify the overlap of themes between exemplars in the causative/inchoative alternation. The degree of overlap would be an indicator of a verb's alternation strength, *i.e.* its capacity to alternate in the two constructions. Such an indicator would allow Lemmens to evaluate the theoretical issue of whether alternations are an essential part of the grammar (in line with generative approaches to language) or whether the alternating constructions should be considered in their own terms with alternations playing only a minor role (in line with Goldberg's (2002) "surface generalization hypothesis").

Lemmens suggests both a quantitative and qualitative account of this overlap through two indices: the Shared Token Index and the Shared Type Index.

The Shared Type Index is a first, rough indicator of the qualitative overlap between members of the same slot in two patterns; in Lemmens' study, these concerns the overlap of themes in two constructions (causative *vs.* inchoative) for a given verb. The Shared Type Index is the ratio of the number of types that occur in both patterns to the total number of types that occur in the two patterns; for two patterns P_1 and P_2:

$$SharedTypeIndex(P_1, P_2) = \frac{NbThemeTypes(P_1 \cap P_2)}{NbThemeTypes(P_1 \cup P_2)}$$

For example, Lemmens (2007) reports that the type frequency of the theme argument of *break* in the British National Corpus is 156 in the causative construction (Cx_1), and 139 in the inchoative construction (Cx_2); the total number of distinct types is 333, and 38 of them occur in both constructions. So, the Shared Type Index between the causative and the inchoative constructions with *break* is $SharedTypeIndex(Cx_1, Cx_2) = \frac{NbThemeTypes(Cx_1 \cap Cx_2)}{NbThemeTypes(Cx_1 \cup Cx_2)} = \frac{38}{333} = 0.114$.

The Shared Token Index is a quantitative index that has to be calculated for each token, that is each verb-theme couple in Lemmens' study. It captures the quantitative weight of a given theme overlap in terms of token frequency. In order to take into account the frequency with which themes are shared, Lemmens introduces the notion of *shared token frequency* ($TokenFreq(T, P_1 \cap P_2)$) that captures the discrepancy between frequencies of a particular token in two distinct patterns; indeed, for a given verb, some themes tend to occur more often with one of the constructions than with the other. In short, Lemmens takes the *shared token frequency* to be equal to the lowest token frequency between the two patterns, times two: $TokenFreq(V, P_1 \cap P_2) = Min(TokenFreq(V, P_1), TokenFreq(V, P_2)) \times 2$ (for a complete illustrated explanation see Lemmens (2007:40–45)). So, for a given theme T that occurs both in patterns P_1 and P_2, the Shared Type Index is the ratio of the shared token frequency to the overall token frequency in both patterns, *i.e.*:

$$SharedTokenIndex(T, P_1, P_2) = \frac{TokenFreq(T, P_1 \cap P_2)}{TokenFreq(T, P_1 \cup P_2)}$$

For example, the theme *news* occured 16 times with break in the causative construction, and 5 times with the inchoative construction. The shared token frequency is equal to the lowest frequency times 2, *i.e.* $TokenFreq(news, Cx_1 \cap Cx_2) = Min(16, 5) \times 2 = 10$. The Shared Token Index between the causative and the inchoative constructions with *break+news* is $SharedTokenIndex(news, P_1, P_2) = \frac{TokenFreq(news, P_1 \cap P_2)}{TokenFreq(news, P_1 \cup P_2)} = \frac{10}{16+5} = 0.476$.

As we said, Lemmens' indices are used to capture the alternation strength and as can be seen with these examples, it is done so by measuring the overlap in alternating arguments. We suggest that these indices can be adapted to our purposes to compare the verbal distribution of two patterns by quantifying the overlap between them. They should capture the extent to which verbs occur in a given pattern or another, in the same way as they capture the extent to which themes occur in both members of an alternation.

In this study, the Shared Type Index will capture the similarity between the distributions of two patterns in terms of verb types (instead of themes). So for two patterns P_1 and P_2, the Shared Type Index is the ratio of the number of verb types that occur both in P_1 and P_2 to the total number of distinct verb types in P_1 and P_2:

$$SharedTypeIndex(P_1, P_2) = \frac{NbVerbTypes(P_1 \cap P_2)}{NbVerbTypes(P_1 \cup P_2)}$$

In the subsequent tests, P_1 will be a functional pattern F constrained by some preposition and P_2 will be either the unconstrained functional pattern F (with all other prepositions) or the functional pat-

P_1/P_2	Distribution attraction	Shared Type Index	Average Shared Token Index
at/to	1.295	0.025	0.076
at/into	0.8	0	0
at/from	0.762	0.037	0.074
at/through	0.504	0.059	0.083

Table 4.8: Schematicity tests for *Subj-V-Obl* and *at vs.* directional prepositions

tern F constrained by another specific preposition (following the two perspectives previously introduced).

The Shared Token Index will also be used to capture the similarity between the verbal distributions of two patterns, but in addition to the account of Shared Type Index, it will be balanced with the statistical weight of each verb type in terms of token frequency. For a given verb V, the Shared Type Index is the ratio of the shared token frequency to the overall token frequency in both patterns, *i.e.*:

$$SharedTokenIndex(V, P_1, P_2) = \frac{TokenFreq(V,P_1\cap P_2)}{TokenFreq(T,P_1\cup P_2)}$$

As previously, the shared token frequency (numerator) is equal to lowest token frequency between the two patterns times two ($TokenFreq(V, P_1 \cap P_2) = Min(TokenFreq(V, P_1), TokenFreq(V, P_2)) \times 2$). Note that it can be equal to 0 if the verb does not occur in at least one of the patterns.

The Shared Token Index is calculated for one given verb and two patterns; what we want to compare is the distribution of two patterns. Thus, for all the verb types occuring in the two patterns, we will compute the Shared Token Index, and the resulting value characterizing the distributional similarity will be the average value for all verbs:

$$SharedTokenIndex(P_1, P_2) = \frac{\sum SharedTokenIndex(V\in P_1\cup P_2, P_1, P_2)}{NbVerbTypes(P_1\cup P_2)}$$

4.5.3 Evaluation

Only one of the constructions in our samples features a formal constraint: the conative construction. The calculation thus concerns the quantification of the similarity between the distribution of the conative construction (the functional pattern [Subj-V-Obl] constrained by *at*) and the intransitive motion construction (the functional pattern [Subj-V-Obl] constrained by a wide range of prepositions expressing a path). To do so, we compute the similarity between the distribution of the preposition *at* and that of other prepositions typical of the intransitive motion construction. The calculation has been limited to the four most frequent prepositions in the IM construction, which nevertheless account for 152 out of 232 instances of the construction: *to* (63 tokens), *into* (36 tokens), *from* (31 tokens) and *through* (22 tokens). The results are summed up in table 4.8.

As expected, the figures for the Shared Type Index and the average Shared Token Index are very close to 0 in all four cases, which is very satisfactory because it suggests that the patterns [Subj-V-Obl$_{at}$] and [Subj-V-Obl$_{to/into/from/through}$] recruit their verbs in different sets, both qualitatively and quantitatively. Nevertheless, none of the distribution attraction values reach the theoretical threshold of 1.3013, *i.e.* the 5% chance level (the *at*/*to* ratio comes very close though). This is quite unexpected, but as we will see, it is not that bad. It should be interesting though to look more closely at the figures for each verb in the distribution of *at*, in order to check whether the figures in table 4.8 reflect a general tendency or whether some tokens in particular are responsible for a discrepancy. The average attraction coefficient for each verb in the distribution of *at* towards the prepositions *to*, *into*, *from* and *through* are presented in table 4.9. What is striking is, first, that most verbs in the distribution of *at* only occur once with this preposition; secondly, the motion verb *run* that occurs in the pattern but is not actually licensed by the conative construction (but by the intransitive motion construction), disprefers *at* compared to *from* and

Verb	Frequency with *at*	*to*	*into*	*from*	*through*
look	13	8.643	6.328	6.539	4.578
aim	2	1.581	0.857	0.887	0.609
stare	2	1.581	0.857	0.887	0.609
nag	1	0.785	0.424	0.439	0.301
scoff	1	0.785	0.424	0.439	0.301
scream	1	0.785	0.424	0.439	0.301
curse	1	0.785	0.424	0.439	0.301
nod	1	0.785	0.424	0.439	0.301
peer	1	0.785	0.424	-0.154	0.301
fire	1	0.785	0.424	0.439	0.301
lunge	1	0.785	0.424	0.439	0.301
beat	1	0.785	0.424	0.439	0.301
glance	1	0.785	0.424	0.439	-0.123
run	1	0.520	0.424	-0.154	-0.514
hint	1	0.785	0.424	0.439	0.301
dive	1	0.785	0.424	0.439	0.301
scrub	1	0.785	0.424	0.439	0.301
sneer	1	0.785	0.424	0.439	0.301

Table 4.9: Attraction coefficient in *Subj-V-Obl* towards *at vs.* directional prepositions

P_1/P_2	Distribution attraction	Shared Type Index	Average Shared Token Index
to/into	0.198	0.089	0.106
to/from	0.219	0.109	0.114
to/through	0.167	0.038	0.031
into/from	0.373	0.108	0.183
into/through	0.158	0.083	0.078
from/through	0.134	0.057	0.052

Table 4.10: Schematicity tests for *Subj-V-Obl* between directional prepositions

through, though not very strikingly. The only verbs that corroborate the intuition that the [Subj-V-Obl_{at}] pattern selects very specific verbs that could be taken as evidence of constructional behavior, are *look* (13 tokens), *aim* and *stare* (2 tokens). These poor results should not be surprising with frequencies of 1: the resulting figures do not mean that those verbs occur with all prepositions, but a single occurence with *at* is no sufficient ground to evaluate whether the verb is attracted by the *at*-pattern, according to the Fisher exact test. The statistical evidence is just not striking enough. Then it seems that for such cases with low frequencies, the shared token/type indices are more effective than collostructional analysis, which essentially relies on the frequencies of each token. We should be careful, however, about the fact that the shared token/type indices do not take statistical significance into account (as collostructional analysis does).

So far we tested to what extent two distributions differ according to a formal constraint that distinguishes two constructions; our indices should account for the fact that patterns licensed by the same constructions have similar distributions. To conclude this battery of tests, we suggest to test whether the indices capture the internal coherence of the intransitive motion construction, *i.e.* whether different prepositions occuring in this construction have similar distributions, which is one of the hypothesis underlying these indices. To do so, we are going to compute these indices to compare the distributions of the four most frequent prepositions in the intransitive motion construction to one another. The results are reported in table 4.10.

Verb	Attraction	Shared Token Index	Token Frequency
come	-0.494	0.609	30
go	0.231	0.476	24
move	-1.849	0	13
pass	-0.61	0	9
run	0.122	0	9
get	-0.816	0.889	9
crawl	-0.365	1	7
turn	0.898	0.154	7
return	0.744	0	6
walk	0.493	0	5

Table 4.11: Schematicity tests for *Subj-V-Obl* and *to vs. into*

The figures for the distribution attraction are quite low as expected, a little lower than when the distribution of the same prepositions was compared to that of *at* (table 4.8). The figures for the shared type/token indices however, are not what we expected: even though they indeed are a bit higher than in table 4.8, they should not be that close to 0; we actually expected to find results close to 1, which would indicate a high similarity between the verbal distributions and hence, according to our hypothesis, a constant meaning of the pattern. Such results, and the shared type index in particular, indicate that, on the contrary, the four prepositions seem to co-occur with quite different verb types. To check this out, we took the ten most frequent verbs of the ditransitive, and for each of these, we computed, for the *to*/*into* couple (the two most frequent prepositions of the intransitive motion) the individual attraction coefficient and shared token index which gives a quantitative account of the distribution overlap. The results are reported in table 4.11.

The results clearly show that *to* and *into* tend to occur with different verbs in the pattern [Subj-V-Obl]. The particularly strong attraction of *move* (significant at the level of almost 1%) to *into* is noteworthy, because it would not be expected to be so for a verb with such a general semantics; *move* can certainly be argued to be the most general verb of motion, even though this is not the most frequent one here. All of the other verbs stay in the range of $[-1.3013, 1.3013]$, *i.e.* outside the theoretical threshold range for attraction or repulsion; but none is as close to 0 as one could expect, and the figures are very variable. The conclusion is even more striking with the shared token indices: half the values are zeros, which clearly shows that those verbs appear with one preposition and not the other. The only results that could be considered satisfactory are those of *come*, *go*, *get* and *crawl.* Very strangely for such a specific verb, *crawl* displays a perfect overlap (1) between the two prepositions. All these results contribute to dramatically weaken the hypothesis on which these indices relie, *i.e.* that all formal variants of a given construction share similar distributions, and that the contrary would be evidence that a formal constraint is a distinctive feature between two constructions. However, it should pointed out that this hypothesis entails a mere statistical tendency and thus in turn relies on the assumption that the sample under study is large enough; in other words, this assumption predicts that the bigger the language sample under study, the more this tendency should be salient. The size of the corpus could again be taken responsible, so these results should not discourage us from trying the same procedure on a bigger corpus that would be more representative of the language.

In sum, the schematicity indices gave results that could allow to discriminate between constructions with the same functional pattern on the basis of their verbal distribution, though not as strikingly as we expected. The results do not allow to decide whether the hypothesis underlying these indices, *i.e.* that constructional meaning can be approximated through verbal distribution, is wrong, or whether the indices have been inappropriatly designed. However, it seems again that the size of the corpus might be responsible for such inconclusive results, especially the small sample we used (only 27 instances of the conative construction). Since these indices are based on distributional information, they are indeed all the more sensitive to the quantitative parameters. It will then be necessary to test this method on a bigger corpus.

Chapter 5

Conclusion and prospects

5.1 Summary and evaluation

This study is an attempt towards the integration of corpus linguistics and cognitive linguistics. It paves the way to a program whose ultimate goal is to automatically discover argument structure constructions in a corpus. To do so, we adopt the hypothesis suggested by Stefanowitsch (2006) that grammatical facts can be derived from corpus data, provided the right method is adopted. This work further builds on other attempts at reconciliating corpora and grammar, such as collostructional analysis (Stefanowitsch and Gries 2003, Gries and Stefanowitsch 2004a;b).

The starting point for our work is Goldberg's model of argument structure constructions developed in (Goldberg 1995; 2005), in which we looked for distinctive features of ASCs that could be observed in a corpus. From this model we derive the assumption that the first level of distinction between ASCs is captured in terms of grammatical functions. It was emphasized, however, that all grammatical functions are not contributed by an ASC, which means in turn that all surface functional patterns do not correspond to an ASC. Grammatical functions in the Susanne corpus were sorted in three groups:

- direct grammatical relations (subject, direct object, second object of ditransitive), considered to be always contributed by an ASC;
- adjunct functions (like time and manner) that were filtered out of the sample, since they can never be contributed by an ASC;
- a third group of functions with an unclear status: they may correspond to arguments of an ASC: oblique, predicatives, etc.

We argued that a first step towards the identification of ASCs in a corpus would be to find quantitative criteria that could provide evidence that a given phrase is indeed an argument (as opposed to what we termed a semi-adjunct). However, we also provided evidence that characterizing ASCs in terms of functions was not enough to distinguish them. Functional patterns are subject to homonymy, in other words, a given functional pattern can correspond to several distinct ASCs. The distinction between homonymic patterns can be captured by accounting for selectional constraints. We emphasized formal constraints on constructions, since they are the only feature directly accessible in a corpus, but we also showed the importance of semantics in some cases; however we suggested that statistical tendencies should overcome this issue and we decided to check as far as we could go with an exclusively form-based approach. We argued that the second step of ASC identification was the quantification of features that could discriminate between functionally equivalent constructions. We decided to limit the study to the case of oblique phrases.

We designed and evaluated four sets of indices to quantify several properties of ASCs:

- Since argument roles of ASCs can be contributed to the clause even when there is no corresponding participant role of the verb, if the attraction of the verb towards the reduced pattern, *i.e.* the same pattern without the oblique phrase is greater than or equal to 0, this is evidence that the instance is licensed by an ASC with a constructional role.

- According to the skewed input hypothesis, if a given verb-preposition couple occurring in a functional pattern F corresponds to an ASC, then there should exist one verb whose token frequency in this pattern-preposition configuration accounts for a large part of the occurrences. Given the assumed semantic relationship of this verb towards the ASC, the verb-ASC attraction should be very high. On the contrary, patterns with a semi-adjunct should not have such a skewed distribution, *i.e.* the highest frequency verb should not account for much of the occurrences of the pattern and it should not be strongly attracted by the pattern.

- According to Fabre and Frérot (2002), productivity scores of the triplet (verb, preposition, complement) are valid indices for the argument/adjunct distinction. We hypothesized that the indices Fabre and Frérot suggest can be applied to the disambiguation between ASCs and semi-adjuncts patterns, *i.e.* a high type frequency of the governor couple (verb, preposition) and a low type frequency of the governee couple (preposition, complement) should be an indicator that an oblique phrase is an argument of an ASC, whereas a low type frequency of the governor couple (verb, preposition) and a high type frequency of the governee couple should be an indicator that the oblique phrase is a semi-adjunct.

- Assuming that constructional meaning can be captured in a corpus by the verbal distribution of the construction, the similarity between the distributions of two patterns with the same functions but constrained by different prepositions should capture whether or not these patterns are licensed by the same construction, in other words they should capture differences in the schematicity of constructions. First, the average attraction of all verbs in both patterns to each preposition, and second, the shared type and token index taken from Lemmens (2007), are used to evaluate qualitatively and quantitatively the degree of similarity between distributions.

The indices were calculated on the Susanne corpus. We evaluated the behaviour of those indices, both qualitatively and quantitatively on three samples of constructions that we isolated from the corpus. In a nutshell, the results are as follows.

The attraction to the reduced pattern provides evidence for the presence of an ASC for approximately half the instances in both cases of the intransitive motion construction and the caused motion construction. In the threshold interval we chosed, the verb types and the attraction values are in line with the idea that constructional meaning can be described in terms of a prototype, as evidenced by previous experiments (Gries 2003, Gries and Stefanowitsch 2006). Only a marginal number of verbs are inaccurately captured by these indices: they are in the list although their occurence in the reduced pattern cannot be explained by the ASC, but rather, for example, by some pragmatic effects (*i.e.* an argument is left unexpressed because it is salient enough in the context).

Skewed frequencies can be observed in all three constructions and the attraction of the syntactic pattern to the most frequent verb is very high as expected. Again, the results confirm the claims by Goldberg and other construction grammarians about constructional meaning, since they basically are general purpose verbs, except in a few cases. Again, such minor inconsistencies with Goldberg's theory of ASCs can be explained by the limited size of the Susanne corpus. However, the results do not allow to validate the hypotheses that all formal variants have the same MFV and that there is only one underlying ASC per pattern. For a few formal variants, we found skewed frequency for verbs that were actually incompatible with the construction under study. Such cases of MFV are not supposed to support the existence of the construction corresponding to the pattern under study. All in all, it appears that detecting skewed frequencies on the basis of formal and functional cues only cannot reliably provide evidence for the presence of an ASC.

The results we got in trying to use the indices from Fabre and Frérot (2002) to the distinction between arguments of ASCs and semi-adjuncts do not allow us to conclude whether this technique is appropriate. Contrary to the results Fabre and Frérot obtain for the argument/adjunct distinction, the governor productivity was generally low. However, as the governee productivity seems indeed to be severely affected by the size of the corpus, the experiment should be carried out again on a bigger corpus, so as to provide a conclusive result.

The results of the schematicity indices, while not so bad, do not allow us, however, to decide whether they can be used to discriminate between patterns. They cannot tell us either whether the underlying hypothesis is wrong (*i.e.* that constructional meaning can be approximated through verbal distribution) or whether the indices have been inappropriately designed. Besides, they contradict the hypothesis that all formal variants of a construction have the same distribution. However, we strongly suggest that the size of the corpus might again be responsible for such inconclusive results, especially given the small sample we used: there were only 27 instances of the conative construction and many verbs only occured once in that construction, which potentially led to the impoverishment of the statistical significance of the results. Since these indices are based on distributional information, they are indeed all the more sensitive to quantitative parameters.

In short, part of those results lie in line of what is to be expected, but they nevertheless have been shown to be insufficient, because of the various reasons we mentioned. These experiments clearly show the limits of a quantitative approach exclusively based on formal cues. In the next and final section of this study, we mention some of the issues of our approach and we suggest possible remedies that could be developed in subsequent work. As a conclusion, we give further ideas and some criticisms about Goldberg's model, on which our whole approach was based.

5.2 Issues, remedy and prospects

5.2.1 The size issue

Throughout the summary of the evaluation of the indices, we mentioned several times that the size of the corpus might be responsible for the poverty of some results, and that using a bigger corpus would significantly improve them. This is particularly true of three measures, because of their reliance on a large amount of quantitative data:

- We hypothesized that all formal variants of a given ASC should have the same most frequent verb. For example, the oblique phrase of the intransitive motion construction denotes a path, so it should not be surprising that the most frequent verb of each formal variant would be a basic purpose motion verb. We further hypothesize that all formal variants should have the *same* most frequent verb, but Susanne did not really confirm this hypothesis;

- The indices we took from Fabre and Frérot (2002), based on two productivity measures involving the triplet (verb, preposition, complement), are also sensitive to the size of the corpus. It should be pointed out again, however, that Fabre and Frérot's experiment was carried out on a corpus from a specific register (geomorphology) and it is unclear how these indices should behave on a general register corpus;

- The schematicity indices are based on distributional information; as we said, such information are very sensitive to the size of the corpus. When a general tendency in a distribution is meant to reflect a semantic property, increasing the amount of data improves the quality of the results.

Testing these indices again on a bigger corpus is then a priority. Note that, as we said in Chapter 2, since our approach relies on grammatical functions and constituency, it basically needs a syntactically annotated corpus (a treebank), but such corpora are limited by their size, since they crucially need manual annotation or checking. An easy way out to get a very large corpus would be to find possible ways through which we could use data with a lower level of annotation (ideally, raw text, more probably PoS tagged).

5.2.2 The lack of intuitive criteria

In Chapter 4, we pointed out a methodological problem that proved to be a major stumbling block for the evaluation of the indices. What we wanted to evaluate was whether the statistical indices we suggested could be used to provide evidence of ASC, *i.e.* if they could be useful in the disambiguation between argument roles and semi-adjuncts. In short, we argued at this point that it is not yet possible to perform such evaluation because of two problems. The first problem is that there is currently no true standard, *i.e.* a generally acknowledged list of ASCs we could evaluate our results against, since this is precisely one of the goals underlying the automatic acquisition of ASCs in corpora. Another possiblity was then to manually evaluate the results, *i.e.* to confront them to the judgement of a speaker. Gries et al. (2005) suggest as a criterion that "for an expression to qualify as a construction, [...] it cannot be compositionally derived in both *form* and/or *meaning* from other constructions available in the language" (p. 639). The second problem is then that deciding whether a given pattern is an ASC or not is a very complex task, that cannot only relie of the intuition of the speaker.

The solution we suggested to overcome these issues is to limit the study to three well-known constructions. But such a decision also limits the scope of our results, which are theoretically valid only for those three constructions. The lack of a simple and objective criteria to identify ASCs is a major limitation to our approach, with respect to both linguistic theory and computational applications. It is very likely that any empirical studies on ASCs would crucially need to address such a question and reconsider Goldberg's model.

5.2.3 Reintegrating semantics

As we said, the approach we adopted was essentially based on grammatical functions and form. But we should emphasize the fact that constructions are form-meaning pairs, which means that both members of the pair are involved in the identification of constructions.

The results of the skewed frequencies experiment show that several meanings can correspond to a given formal variant of a functional pattern, in the sense that it can accommodate very different verbs with very different semantics: for example, the pattern [Subj-V-Obl$_{for}$] can be used to encode a motion meaning, as in *Summers headed for the street*, but not exclusively; the sentences *Former Vice-President Richard M. Nixon called for a firmer and tougher policy toward the Soviet Union* and *he would ask for exclusive use of voting machines in the state by January, 1964* do not denote a motion.

In addition, there also exist strong collocations of a verb and a preposition whose meaning is very conventionalized; for example, the couples *believe in* and *depend on*. Such strong collocations are a possible explanation to the presence of biased frequencies that do not correspond to an ASC, and more generally they can seriously affect the results of a form-based identification.

The categorization process of ASCs actually involve complex semantic interactions where the meaning of the verb as well as that of each argument play a crucial role. The notion of semantic compatilibility is central to ASCs, as many studies emphasized. It appears that we should find a way to reintegrate semantics into the quantitative analysis.

5.2.4 A critique of Goldberg's model

The whole approach of ASC identification we presented in this study was based on Goldberg's model, and throughout the study we had the opportunity to evaluate its empirical value on corpus data. As a conclusion to this study, we would like to outline some reflections about this model and express some criticism.

We already said a few words about the uncertain status of grammatical functions in Goldberg's theory of argument structure, an analysis we would like to develop here. The main problem of using grammatical functions as the basis of ASCs is that those functions do not have any formal marking in English; this is especially a problem when dealing with corpus data where only form is accessible. Of course, one could argue that some categories can be assigned a function on the basis of form only; for example, personal pronouns such as *I*, *we* or *they* are always subjects; but in most cases, there is no one-to-one mapping between form and grammatical function. If there is no reliable formal marking with minimal ambiguity, one may wonder how grammatical functions are assigned.

We can assume that they are specified at the level of the verb, in other words that they correspond to a lexical level of complement selection that assigns functions to complements of the verbs according to their syntactic position, which in turn are selected by argument structure constructions. This seems to be the most straightforward hypothesis, but it poses one major problem. Argument roles have to be fused with a function provided by the verb; but if it is a contructional role, there is not necessarily a corresponding participant role of the verb: it can be contributed by the verb only. If an ASC with a constructional role is instantiated to contribute an additional argument role to the verb, for example in the case of *bake* in the ditransitive construction, there is by definition no corresponding participant role to be fused with this argument role. A possible solution would be to posit several entries for each verb, each with a different list of functions. But such a solution would actually lead us back to the shortcoming of the lexicalist account that the constructional approach to argument structure is supposed to avoid, *i.e.* the proliferation of lexical entries. In such a view, verbs still have implausible entries, even though they not express in terms of roles but in terms of functions.

We can also consider that functions are categories themselves: assigning functions would correspond to assign a functional category to each clause level phrase, given a couple of formal cues, like word position or case and formal category for example. Word position would be exclusively preferred in English and other constrained word order languages that do not feature case-marking (except of course for pronouns; in addition, case marking could be used in inflecting languages and would be exclusively preferred in free word order languages. Anyway, the question of the status of grammatical functions in Goldberg's theory and of whether the model really needs grammatical functions or whether it could be based on empirically observable cues should be addressed in more detail.

The second question we would like to address is the status of semi-adjuncts compared to arguments and if this difference is really motivated. In terms of relationship to the verb, there is no difference between an ASC with only verbal roles (or any instance of a construction with a perfect isomorphy between participant and argument roles) and a pattern with a semi-adjunct: in both cases, all arguments are selected (and required) by the verb.

According to Goldberg's model, the only difference between the two types of complementation is semantic: the ASC contributes an additional meaning to the clause, whereas the meaning of the semi-adjunct pattern is fully predictable from the meaning of other constructions in the broadest sense: the verb, the prepositional phrase, etc. It should be pointed out, however, that in the case when there is a perfect isomorphy between the participant roles of the verbs and the argument roles of the construction, there is no evidence of an additional meaning provided by the construction itself. More precisely, the meaning of the verb and the meaning of the construction overlap, so in a way the construction is not required in order to construe the meaning of the whole clause: in the example, the meaning of a transfer is already provided by the verb. If all verbs occuring in the ditransitive pattern had a transfer meaning, there would be no need to posit an ASC, and there would actually be no observable evidence that such a construction exists. So, in a way, ASCs are only needed when "unusual" verbs occur in the corresponding functional pattern.

In the example of semi-adjunct pattern we mentioned in Chapter 3, *X load Y with Z*, it can actually be argued that the pattern does convey a transfer meaning that can be generalized over instances. The same pattern can occur with a range of verbs and all cases denote a kind of transfer. The meaning of the verb is only required to specify the way the theme is transfered and its final state, *e.g. load* specifies that the final location theme is inside the container, *spray* denotes that the theme has been spread onto a surface, etc. Such an ASC could be posited in virtue of the fact that all instances of the pattern share a common meaning. The only difference with an attested ASC is that no case has been reported (yet) of a verb with no prior transfer meaning occuring in this pattern, *i.e.* no case have been attested when the transfer meaning would arguably be contributed by the construction alone.

However, a common position in cognitive linguistics is that redundancy can be posited in the language model, *i.e.* both rules and instances can be present in the language model at the same time. Many linguistic theories advocate economy of representation and seek maximally general rules; what Langacker (1987) terms the *rule/list fallacy* lies in the assumption that knowledge of a rule necessarily expunges knowledge of the instances of that rule. The fact that such rules can be formulated and used by speakers does not necessarily entail that speakers always do apply the rules. It is conceivable that they have stored a vast array of complex linguistic forms and that these are accessed, ready-made, when required. Such an account is psychologically relevant and stems from the cognitive reality that we can both compute and store the same information. For example, words in cognitive grammar are independent units and thus have their own meaning, but they can still also be decomposable by morphological rules; for example, *speaker* can be decomposed as the combination of the smaller units *speak* and the agentive suffix *-er*.

In our case, we argue in virtue of this principle that positing such a distinction between semi-adjunct patterns and ASCs is not sufficiently motivated in a cognitive grammar approach. There should be no difference in status between them, since the former can be stored as instances of other constructions, just as the latter have to be stored as independent linguistic units. All patterns of complementation should have the same constructional status.

5.2.5 Perspectives

A possible solution that we suggest in order to solve the issues mentioned in this chapter would be to partly shift the perspective of our approach. The approach we advocated is fundamentally bottom-up, *i.e.* we decomposed clauses into smaller units (phrases bearing a grammatical function) and we derived indices from each individual 'building block' to make general patterns appear. What we suggest instead is to switch to a more holistic account whose aim would be to capture recurring patterns of complementation as a whole. Of course, this implies interpreting Goldberg's model (the starting point of our analysis) more loosely, the theoretical implications of which are considerable. In particular, we strongly suggest to study what British linguistics, and in particular *Pattern Grammar* Hunston and Francis (2000), could bring to our problem. In short, Pattern Grammar is a phraseological, corpus-based approach to grammar; it shares some of the goals of construction grammars while adopting a radically holistic approach.

Another perspective we should explore further is Harris distributional analysis (Harris 1951; 1968). Harris claims that syntactic and semantic relations, expressed in terms of operator and operands, can be detected in a corpus through a rigorous and systematic analysis as presented in Harris (1988; 1991). In the last decade, the distributionalist enterprise has known a revival in Natural Language Processing because one of its major benefits is to eventually reconcile formal linguistics and information theory (cf. Habert and Zweigenbaum 2002). In subsequent work, we should study whether it would possible to adapt this method for the extraction of regular syntactic patterns, as Balvet (2002a;b) suggests. Such a solution would also be computationally interesting since it would not need syntactic annotations.

Bibliography

Baker, C. L. (1979). Syntactic Theory and the Projection Problem. *Linguistic Inquiry 10*(4), 533–581.

Balvet, A. (2002a). *Approches catégoriques et non catégoriques en linguistique des corpus spécialisés, application à un système de filtrage d'information.* Ph. D. thesis, Université Paris X.

Balvet, A. (2002b). LIZARD, un assistant pour le développement de ressources linguistiques à base de cascades de transducteurs. In *Actes de RÉCITAL 2002, Nancy, 24-27 juin 2002.* INRIA.

Barlow, M. and S. Kemmer (Eds.) (2000). *Usage-based models of language.* Stanford, California: CSLI Publications.

Berlin, B. and P. Kay (1969). *Basic Color Terms: Their Universality and Evolution.* Berkeley and Los Angeles: University of California Press.

Birner, B. J. (1994). Information status and word order: An analysis of English inversion. *Language 70*(2), 233–259.

Bourigault, D. and C. Fabre (2000). Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de Grammaire 25*, 131–151.

Bourigault, D. and C. Frérot (2004). Ambiguïté de rattachement prépositionnel : introduction de ressources exogènes de sous-catégorisation dans un analyseur syntaxique de corpus endogène. In *Actes de la XIe Conférence sur le Traitement Automatique des Langues Naturelles, Fès, Maroc, 19-21 avril 2004.*

Bourigault, D. and C. Frérot (2006). Acquisition et évaluation sur corpus de propriétés de sous-catégorisation syntaxique. *Traitement Automatique des Langues 47*(3), 141–154.

Bybee, J. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes 10*(5), 425–455.

Bybee, J. (2003). Mechanisms of change in grammaticization: The role of frequency. In B. D. Joseph and R. D. Janda (Eds.), *The Handbook of Historical Linguistics*, pp. 602–623. Oxford: Blackwell Publishing.

Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language 82*(4), 711–733.

Bybee, J. and S. Thompson (1997). Three frequency effects in syntax. *Berkeley Linguistics Society 23*, 378–388.

Casenhiser, D. and A. E. Goldberg (2005). Fast mapping between a phrasal form and meaning. *Developmental Science 8*(6), 500–508.

Chomsky, N. (1957). *Syntactic Structures.* Walter de Gruyter.

Chomsky, N. (1965a). *Aspects of the Theory of Syntax.* MIT Press.

Chomsky, N. (1965b). Formal discussion: the development of grammar in child language. In U. Bellugi and R. Brown (Eds.), *The Acquisition of Language*, pp. 35–9. Purdue University, Indiana.

Croft, W. (1998). Linguistic evidence and mental representations. *Cognitive Linguistics 9*(2), 151–173.

Croft, W. (2001). *Radical construction grammar: syntactic theory in typological perspective.* Oxford & New York: Oxford University Press.

Croft, W. (2004). Logical and typological arguments for radical construction grammar. In M. Fried and J.-O. Östman (Eds.), *Construction Grammar(s): Cognitive and cross-language dimensions*, Number 3 in Constructional Approaches to Language, pp. 273–314. Amsterdam: John Benjamins.

Dirven, R. and J. R. Taylor (1988). The conceptualization of vertical space in English: the case of" tall. In B. Rudzka-Ostyn (Ed.), *Topics in Cognitive Linguistics*, pp. 379–402. Amsterdam: John Benjamins.

Evans, V., B. K. Bergen, and J. Zinken (2007). The cognitive linguistics enterprise: an overview. In V. Evans, B. K. Bergen, and J. Zinken (Eds.), *The Cognitive Linguistics Reader*, Chapter 1. London: Equinox Publishing.

Fabre, C. and D. Bourigault (2001). Linguistic clues for corpus-based acquisition of lexical dependencies. In *Proceedings of the Corpus Linguistic Conference, Lancaster University, UK.*

Fabre, C. and C. Frérot (2002). Groupes prépositionnels arguments ou circonstants : vers un repérage automatique en corpus. In *Actes de la Conférence TALN, Nancy 24-27 juin*, pp. 215–224.

Fillmore, C. (1977). Topics in lexical semantics. *Current issues in linguistic theory 76*, 76–138.

Fillmore, C. J. (1982). Towards a Descriptive Framework for Spatial Deixis. In R. J. Jarvella and W. Kleib (Eds.), *Speech, Place and Action*, pp. 31–59. London: John Wiley & Sons Ltd.

Fillmore, C. J. (1985). Frames and the semantics of understanding. *Quaderni di Semantica 6*(2), 222–254.

Fillmore, C. J. (1991). "Corpus Linguistics" or "Computer-Aided Armchair Linguistics". In J. Svartvik (Ed.), *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4–8 August 1991*, Berlin/New York, pp. 35–60. Walter de Gruyter.

Fillmore, C. J., P. Kay, and M. C. O'Connor (1988). Regularity and Idiomaticity in Grammatical Constructions: The Case of Let Alone. *Language 64*(3), 501–538.

Goldberg, A. E. (1995). *Constructions: a construction grammar approach to argument structure.* Chicago: University of Chicago Press.

Goldberg, A. E. (2002). Surface generalizations: An alternative to alternations. *Cognitive Linguistics 13*(4), 327–356.

Goldberg, A. E. (2005). *Constructions at Work: The Nature of Generalization in Language.* Oxford: Oxford University Press.

Goldberg, A. E., D. Casenhiser, and N. Sethuraman (2003). A lexically based proposal of argument structure meaning. In *Proceedings of the Annual Chicago Linguistics Society.*

Goldberg, A. E., D. Casenhiser, and N. Sethuraman (2004). Learning argument structure generalizations. *Cognitive Linguistics 15*(3), 289–316.

Goldberg, A. E. and R. Jackendoff (2004). The English resultative as a family of constructions. *Language 80*(3), 532–568.

Gries, S. T. (2003). Towards a corpus-based identification of prototypical instances of constructions. *Annual review of cognitive linguistics 1*, 1–28.

Gries, S. T., B. Hampe, and D. Schönefeld (2005). Converging evidence: bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics 16*(4), 635–76.

Gries, S. T. and A. Stefanowitsch (2004a). Co-varying collexemes in the into-causative. In M. Achard and S. Kemmer (Eds.), *Language, Culture, and Mind*, pp. 225–236. Stanford, California: CSLI.

Gries, S. T. and A. Stefanowitsch (2004b). Extending collostructional analysis. A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics 9*, 97–129.

Gries, S. T. and A. Stefanowitsch (2006). Cluster analysis and the identification of collexeme classes. In J. Newman and S. Rice (Eds.), *Empirical and Experimental Methods in Cognitive/Functional Research*. Stanford, California: CSLI Publications.

Gross, M. (1984). Lexicon-grammar and the syntactic analysis of French. In *Proceedings of the 22nd annual meeting on Association for Computational Linguistics*, pp. 275–282.

Habert, B. and P. Zweigenbaum (2002). Régler les règles. *Traitement automatique des langues 43*(3), 83–105.

Harris, Z. S. (1951). *Structural Linguistics*. University Of Chicago Press.

Harris, Z. S. (1968). *Mathematical structures of language*. New York: Interscience Publishers.

Harris, Z. S. (1988). *Language and information*. New York: Columbia University Press.

Harris, Z. S. (1991). *A Theory of Language and Information: A Mathematical Approach*. Oxford: Clarendon Press.

Hunston, S. and G. Francis (2000). *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English*. John Benjamins.

Jaeger, J. and J. J. Ohala (1984). On the Structure of Phonetic Categories. *Berkeley Linguistic Society 10*, 15–26.

Käding, F. W. (1897). *Häufigkeitswörterbuch der deutschen Sprache*. Steglitz: privately published.

Kay, P. (2002). An informal sketch for a formal architecture for construction grammar. *Grammars 5*, 1–19.

Kay, P. (2005). Argument structure constructions and the argument-adjunct distinction. In M. Fried and H. C. Boas (Eds.), *Grammatical Constructions: Back to the Roots*, pp. 71–98. Amsterdam: Benjamins.

Kay, P. and C. J. Fillmore (1999). Grammatical constructions and linguistic generalizations: the What's X doing Y? construction. *Language 75*, 1–33.

Kleiber, G. (1990). *La sémantique du prototype*. Presses Universitaires de France.

Lakoff, G. (1977). Linguistic Gestalts. In P. S. Beach W.A., Fox S.E. (Ed.), *Papers from the Thirteenth Regional Meeting, Chicago Linguistic Society, April 14-16, 1977*, pp. 236–287.

Lakoff, G. (1987). *Women, fire, and dangerous things: what categories reveal about the mind*. University of Chicago Press.

Lakoff, G. (1990). The invariance hypothesis: Is abstract reason based on image-schemas. *Cognitive Linguistics 1*(1), 39–74.

Lakoff, G. and M. Johnson (1980). *Metaphors We Live by*. University of Chicago Press.

Langacker, R. W. (1987). *Foundations of Cognitive Grammar. Volume 1: Theoretical Prerequisites*. Stanford University Press.

Langacker, R. W. (1991). *Foundations of Cognitive Grammar. Volume 2: Descriptive Application*. Stanford University Press.

Langacker, R. W. (2000). A dynamic usage-based model. In M. Barlow and S. Kemmer (Eds.), *Usage-Based Models of Language*, pp. 1–63. Stanford, California: CSLI Publications.

Leclère, C. (1978). Sur une classe de verbes datifs. *Langue Francaise 39*, 66–75.

Leech, G. (1992). Corpora and Theories of Linguistic Performance. In *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991*, pp. 105–125.

Lemmens, M. (1998). *Lexical perspectives on transitivity and ergativity: causative constructions in English.* Amsterdam, Philadelphia: John Benjamins.

Lemmens, M. (2006). More on objectless transitives and ergativization patterns in english. In D. Schönefeld (Ed.), *Constructions Special Volume 1 – Constructions all over: case studies and theoretical implications.* http://www.constructions-online.de/.

Lemmens, M. (2007). Collostructional analysis of the causative alternation in English. Handout ICLC-10, Krakow, Poland, 15-21 July.

Levin, B. (1993). *English Verb Classes and Alternations : A Preliminary Investigation.* University Of Chicago Press.

Levin, B. (1999). Objecthood: An event structure perspective. *Proceedings of CLS 35*, 223–247.

Levin, B. and M. H. Rappaport (1995). *Unaccusativity: At the Syntax-Lexical Semantics Interface.*, Volume 26 of *Linguistic Inquiry Monograph.* Cambridge: The MIT Press.

McEnery, T. and A. Wilson (2001). *Corpus Linguistics.* Edinburgh University Press.

Meyers, A., C. MacLeod, and R. Grishman (1996). Standardization of the complement/adjunct distinction. In *Proceedings of the Seventh EURALEX International Conference, Göteborg, Sweden.*

Miller, P. (1997). Compléments et circonstants : distinction syntaxique ou sémantique ? *Cycnos 15, n° spécial*, 91–103.

Pedersen, T. (1996). Fishing for exactness. In *Proceedings of the South Central SAS Users Group Conference*, pp. 188–200.

Pollard, C. J. and I. A. Sag (1987). *Information-based syntax and semantics.* CSLI Publications.

Pollard, C. J. and I. A. Sag (1994). *Head-Driven Phrase Structure Grammar.* University Of Chicago Press.

Pulman, S. G. (1983). *Word Meaning and Belief.* London: Croom Helm.

Quirk, R., S. Greenbaum, G. Leech, and J. Svartvik (1972). *A Grammar of Contemporary English.* London and New York: Oxford University Press.

Quirk, R., S. Greenbaum, G. Leech, and J. Svartvik (1985). *A Comprehensive Grammar of the English Language.* London and New York: Longman.

Rappaport, M. H. and B. Levin (1985). A case study in lexical analysis: The locative alternation. Unpublished manuscript, MIT Center for Cognitive Science, Cambridge, MA.

Rice, S. (1988). Unlikely lexical entries. *Berkeley Linguistics Society 14*, 202–212.

Rosch, E. (1973). Natural categories. *Cognitive Psychology 4*(3), 328–350.

Rosch, E. (1978). Principles of categorization. In E. Rosch and B. Lloyd (Eds.), *Cognition and Categorization*, pp. 27–48. Hillsdale: Laurence Erlbaum Associates.

Sampson, G. (1995). *English for the Computer: The SUSANNE Corpus and Analytic Scheme.* Oxford University Press.

Schulte Im Walde, S. (2006). Experiments on the Automatic Induction of German Semantic Verb Classes. *Computational Linguistics 32*(2), 159–194.

Searle, J. R. (1969). *Speech Acts: An Essay in the Philosophy of Language.* Cambridge University Press.

Sinclair, J. (1991). *Corpus, concordance, collocation.* Oxford: Oxford University Press.

Sinclair, J. (1996). The search for units of meaning. *Textus 9*(1), 75–106.

Stefanowitsch, A. (2006). Negative evidence and the raw frequency fallacy. *Corpus Linguistics and Linguistic Theory 2*(1), 61–77.

Stefanowitsch, A. and S. T. Gries (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics 8*(2), 209–243.

Stefanowitsch, A. and S. T. Gries (2005). Covarying collexemes. *Corpus Linguistic and Linguistic Theory 1*(1), 1–43.

Stubbs, M. (1995). Corpus evidence for norms of lexical collocation. In G. Cook and B. Seidlhofer (Eds.), *Principle and practice in Applied Linguistics, Studies in Honour of H.G. Widdowson*, pp. 245–256. Oxford University Press.

Stubbs, M. (2001). *Words and Phrases: Corpus Studies of Lexical Semantics.* Blackwell Publishing.

Taylor, J. R. (1995). *Linguistic categorization.* Oxford University Press.

Vandeloise, C. (1986). *L'espace en français: sémantique des prépositions spatiales.* Paris: Seuil.

Vendler, Z. (1967). *Linguistics in Philosophy.* Cornell University Press.