Gunnel Källgren
Dep of Linguistics
Stockholm University
S-106 91 Stockholm, Sweden

What good is Syntactic Information in the Lexicon of a Syntactic Parser?

Imagine a situation where we want to parse texts and get as output information about the traditional grammatical categories of sentences, i. e. subject, direct and indirect object, finite verb, adverbials etc. That is an explicit goal of most parsers, and a necessary prerequisite for many applications of computational linguistics. How much information, syntactic and/or other, do we need in order to reach that goal?

In this paper I will look at the kind of syntactic information that is mostly given in syntactic parsers (sect. 1) and relate it to different parts of the parsing process (sect. 2). In section 3, I present an algorithm for assigning syntactic function to constituents without any use of lexical look-up. I discuss its results (sect. 4) and the type of errors that arise (sect. 5), which leads me to the conclusion that parsing can in most cases very well be done without a lexicon containing syntactic information. In section 6 I point to the need for different parsers for different purposes.

1. What kind of information can be given?

Many parsing systems use information that is centered around the verb. In the dictionary of the parser, each verb has information about whether it is transitive or intransitive and how obligatory or optional its complements are. Often some kind of selectional restrictions are also given. They describe properties of the referents of the noun phrases that enter different syntactical roles in relation to the verb. Such properties are concrete/abstract, animate/inanimate, human/non-human etc. The apparatus demands marking of the verbs with their functional frames, and marking of all words that can enter those frames accordingly, and procedures for checking the different kinds of marking against each other.

The basic thoughts behind this technique are clear; to get correct syntactic assignments in 1a-b and 2a-b one will need verbal frames like 3 and 4.

(1a) Mary gave the book to John.
(b) Mary gave John the book.
(2a) She drank the wine from a chrystal glass.
(b) The horse drank (the water) (in the bucket).

(3) give(x,y,z) where x=subj, y=dir obj, z=indir obj
patterns: x give y PREP z where PREP=to
x give z y

(4) drink(x,y,z) where x=subj, y=dir obj, z=loc advl
pattern: x drink (y) (PREP z) where PREP=from, in
selectional restrictions: x=animate, y=liquid, z=some kind of container

What good is Syntactic Information in the Lexicon of a Syntactic Parser?
Gunnel Källgren
Proceedings of NODALIDA 1987, pages 5-16

The classificatory system can then be developed and refined to cover more and more of the ways a word is used, and more and more of the vocabulary of a language (see Källgren 1986). Some existing systems use a much more finely graded classification, and some also include e g a classification of the adverbials that can modify a verb or a sentence. Fig. 5 shows two entries from a such system, the Janus system, that seems to be carefully planned and built on a larger scale than most others, see Cumming 1986 for a closer description of its design.

(5)

```
 :name 'LEAD
 :spelling "lead"
 :features '(VERB INFLECTABLE LEXICAL NOT-CASEPREPOSITIONS
OBJECTPERMITTED NOT-TOCOMP NOT-QUESTIONCOMP NOT-PARTICIPLECOMP
NOT-MAKECOMP NOT-BAREINFINITIVECOMP NOT-COPULA PASSIVE NOT-THATCOMP
NOT-ADJECTIVECOMP NONE-OF-BITRANSITIVE-INDIRECTOBJECT DOVERB DISPOSAL
EFFECTIVE OBJECTNOTREQUIRED NOT-OBJECTNOTPERMITTED SUBJECTCOMP
UNITARYSPELLING S-IRR PASTFORM EDPARTICIPLEFORM)
 :properties '((PASTFORM "led") (EDPARTICIPLEFORM "led"))

 :name 'SIMPLE
 :spelling "simple"
 :features '(ADJECTIVE NOT-CASEPREPOSITIONS R-ST DEGREE
COMPLEMENTPERMITTED TOCOMP FORNPPERMITTED SUBJECTHOLD NOT-SUBJECTCOMP
NONE-OF-APPROPRIATENESS-POSSIBILITYPROPERTY-OBVIOUSNESS NOT-THATCOMP
NOT-PREDICATEONLY)
 :properties '()
```

The lexicon must mainly be built up by hand and that is a task that takes good and clear linguistic intuitions to do. Actually, it is quite hard to imagine all the different kinds of complements that a verb can take. Sometimes subtle shades of meaning in the verb accompanies differences in the functional frame ('to read a book' but 'to read a child to sleep'). The verb can then be entered as two homonymous forms, or some other way of entering the information can be designed, but for the parsing process as such this does not necessarily remove all ambiguity. There is always the risk that some possibility is forgotten and the functional frame gets too small, or that some allowed pattern makes the frame too loose and unwanted constructions become allowed. In short, this way of putting conditions on the parse is hard, tedious and time-consuming without ever guaranteeing total correctness.

As a small excercise, the reader might want to try to figure out how the frame of the verb 'to read' should be described, given the corpus in 6, which is only a small subpart of the possible readings (!) of that single verb. What are objects and what are adverbials? When is a preposition not a preposition but an aspect marker? How can we capture that, in our culture, messages can be written on the side of buses, but not on trains or airplanes? And where precisely goes the borderline between signs that can be read and signs that can be read on?

(6a) She read the book.
(b) She read on the train.
(c) She read on the bus that there was a sale on. (Ambiguous.)
(d) She read the street sign to see where she was.
(e) *She read on the street sign to see where she was.
(f) She read on the sign that the store was closing at 6.
(g) *She read on the street sign that the street was 5th Ave.
(h) She read on without being disturbed by the noise.
(i) She read quickly.
(j) The book read quickly.
(k) She read the book quickly.
(l) The thermometer read 20 degrees.
(m) She read 20 degrees on the thermometer.
(n) Every morning she read the thermometer to see what to wear.
(o) She read the child to sleep.
(p) She read the book to pass the exam.
(q) She (read) (up on the mountain).
(r) She (read up) (on the mountain).
(s) She read German.
(t) She read law.

Some systems use an on-line ordinary dictionary instead. This makes large-scale parsing possible to an extent that will probably never be reached with the more handicraft based systems, but it brings with it the problem of extracting the wanted information from dictionary entries that were not written with this particular application in mind. The problem with full coverage of all possibilities also remains, but still this seems to be a more promising way. (Jensen and Binot 1987.)

2. Category assignment versus functional assignment

Parsing really consists of (at least) two parts, corresponding to the important difference between category and function. It is one task to identify the constituents of a sentence and assign their category, another task to decide the functional roles that the constituents play and to assign the structure of the whole sentence. Different languages signal categorial and functional information in different ways and to a different extent, a fact that must influence the mode of parsing chosen for a language. In a language like Finnish, both grammatical category and syntactic function can often be seen from the morphology of single words (Karlsson 1985), whereas in English, knowledge about overall sentence structure is often needed in order to assign the correct category to ambigous words. In 7, the two parsing levels clearly presuppose each other:

(7a) He judges sentence (7b) to be correct.
(b) Judges sentence criminals.

Swedish lies somewhere between Finnish and English in this respect. The morphology gives a good idea of the category of a word, especially when its close context is taken into account (Källgren 1984), while it is mainly the word order of the full sentence that gives information about the syntactic function of its constituents.

If category assignment and functional assignment are kept apart, what kind of information is needed for each of the two tasks? To identify constituents and their category, it is enough to know what

part of speech each word is and the allowed internal patterns of constituents such as noun phrases, prepositional phrases, and simple and complex verbs. If for the moment we disregard the problem of homonymy, we can generally say that a sequence of article + adjective + noun constitutes a noun phrase without knowing about the abstractness or animacy of the noun, and a finite verb is a finite verb, no matter what its functional frame may contain.

For the task of category assignment then, no syntactic information except part of speech and internal structure of certain phrasal categories seems to be needed. All the elaborate apparatus described above must thus be introduced in order to handle the assignment of syntactic function. This seems reasonable; the kind of conditions expressed in verb frames are conditions on sentence structure, not on constituent-internal structure. The purpose of adding the verb frames and the syntactico-semantic properties of the nouns is then to aid in the identification of subject, object, main verb etc. The question will thus be: Precisely what kind of information and how much information is needed in order to reach a good parse?

Tagging and parsing systems for English that make precisely this separation of tasks have been constructed for the large-scale analysis of the Brown and LOB corpora. Their systems for word-tagging, i e about the same as what was above called category assignment, are based on lexicon and morphology. TAGGIT is the system used for the Brown corpus. Its degree of correctness in deciding the grammatical class of words in unrestricted text is reported to be 77% (Greene and Rubin 1971). For CLAWS, the LOB system which is developed on the basis of TAGGIT, the degree of correctness is 96-97% (Garside 1987).

Independently of their work, I have built up a morphologically based system for category assignment in unrestricted Swedish text. For theoretical reasons, I have tried to limit the lexicon as much as possible, and today the system has a lexicon of less than 300 words that belong to closed categories or are morphologically highly irregular, which should be compared to the lexicon of 7 200 words in the CLAWS system. Except for those 300 words, all word-tagging in the Swedish system is based on morphological and contextual clues. As stated above, Swedish morphology contains more information and less ambiguous information than is the case with English, so the performance of the lexicon-less Swedish system is around 90% on any arbitrary text. The system is described in Källgren 1984a,b, 1985. It is written in BETA (see Brodda 1987) and runs on DEC-10/20 computers under TOPS-10/20 and is being implemented on PCs under DOS. Its output can give sentences that are analyzed like 8 and 9, or that simply look like the corresponding structures 10 and 11.

(8) S:(NP:(Mannen) Fin-V:(skrev) NP:(ett brev))
the man wrote a letter

(9) S:(NP:(Kvinnan) Fin-V:(visste) S:(Conj:(att) Pron-subj:(han)
the woman knew that he
Fin-V:(skrev) NP:(ett brev) Advl:(idag)))
wrote a letter today

(10) S:(NP Fin-V NP)
(11) S:(NP Fin-V S:(Conj Pron-subj Fin-V NP Advl))

3. An algorithm for functional assignment in Swedish sentences

To reach a full parse, I have constructed an algorithm for functional assignment in Swedish sentences, which is based on the output from the morphological system. The algorithm is now being implemented in BETA and CommonLisp. Its task is to decide, on the basis of as little information as in 10 and 11, the major syntactic roles of the constituents. I will here present the algorithm and its results and discuss the implications of it.

The algorithm has been applied to natural Swedish texts that have already been analyzed into sentences and clauses, and with the following major categories identified: noun phrases, pronouns in subject and object form, prepositional phrases, finite and non-finite verbs, auxiliaries, adverbs, prepositions, and conjunctions. As mentioned above, a system that does precisely this category assignment with a high degree of correctness (around 90%) already exists, but to be able to judge the output from the different parts of the parse separately, we have started from an idealized situation where the category assignment is taken to be 100% correct. Note that the algorithm does not presuppose correct text, it presupposes text that does not need to be corrected. (Cf. the discussion in section 6 about the different possible purposes of parsers.) This means that it gives a best-possible parse for every sentence, regardless of whether the sentence adheres to grammatical standards or not. A robustness of this kind is necessary to be able to deal with unrestricted text.

We regard each occurrence of a verb and its complements as a simplex sentence. The present version of the algorithm only works with clauses that contain a finite verb. Non-finite clauses pose an extra set of difficulties to any parser. Subordinate clauses get their analysis both as to the functional role they play in the superordinate sentence and as to their internal structure. The basic task is then to identify the subject of each simplex sentence and its direct and indirect object, in case any object(s) occur.

The algorithm is formulated as a set of rules (around 25 at present), of which some describe clear and unequivocal patterns and others give heuristic solutions for situations that can be ambiguous. Some examples:

A noun phrase "inside" a complex verb, i e between a finite and an infite verb, must always be the subject of the verb. Examples like 12a-b thus gives the rule 13.

(12a) Idag har mannen skrivit brevet.
today has the man written the letter
(b) Brevet har mannen skrivit idag.
the letter has the man written today

(13) X Fin-V NP Infin-V Y -> NP := subj

If the position before the finite verb (the so-called fundament position) is filled by an adverbial (an adverb or a prepositional phrase), the first noun phrase after the finite verb must be the subject. Examples 14a-b lead to rule 15.

(14a) Idag skrev mannen brevet.
today wrote the man the letter
(b) I arbetsrummet skrev mannen brevet.
in the study wrote the man the letter

(15a) Advl Fin-V NP X -> NP := subj
(b) PP Fin-V NP X -> NP := subj

Where no such rules are applicable, the heuristic simply says that the first noun phrase in a sentence is its subject. Rule 16 gives a correct result in sentences like 17a but not in 17b.

(16) NP Fin-V X -> NP := subj

(17a) Mannen skriver alla brev i arbetsrummet.
the man writes all letters in the study
(b) Alla brev skriver mannen i arbetsrummet.
all letters writes the man in the study

A set of analogous rules decide the assignment of direct object (mostly simply the noun phrase that is not subject) and the choice between direct and indirect object where two objects occur.

4. Some results of the application of the algorithm

The table below (Fig 18) gives the results for a corpus of 1,451 simplex sentences, taken from different text types. Of the 1,451 simplex sentences, 160 lack a finite verb, so the figures in the table are computed on the basis of the 1,291 sentences that the algorithm applies to.

Instances where the algorithm (correctly) predicts zero subject, as in imperatives, count as a correct identification of subject. For the identification of objects, two figures are given. The first covers the cases where object(s) occur without a preposition in the sentence, while the second also includes instances with particle verbs ('skriva ner något' write down something) and instances where a prepositional phrase can on semantic grounds be regarded as a direct or indirect object. The last category can be difficult to judge, but some clear instances exist ('titta på TV' look at TV, 'ge boken till flickan' give the book to the girl, as compared to 'ge flickan boken' give the girl the book). Generally, a human linguist in many cases has quite a hard time in choosing between particle + object/prepositional object/prepositional adverbial phrase. There are often no clear rules-of-thumb and different people come to different decisions, so the resulting figures in themselves are not that important. The important thing is that such instances occur, and that a weak point of the algorithm is its inability to decide the syntactic function of prepositional phrases. They are at present all regarded as some unspecified kind of adverbials. However, this deficiency is not as destructive as one might fear, since there are in fact not that many such occurrences.

(18) Ratio of correct functional assignments

	%	N
Correct subject/total number of sentences	99,5	1 284/1 291
Correct dir object/"naked" dir objects	98,6	579/587
Correct dir obj/"naked" + prep dir obj	83,5	579/693
Correct indir obj/"naked" indir obj	100	3/3
Correct indir obj/"naked" + prep indir obj	60	3/5

The ratios for identification of subject and direct and indirect object without preposition are remarkably high. Indirect objects are surprisingly infrequent. In the whole material of almost 1,300 sentences only five indirect objects appear, three without a preposition that are all correctly identified and two instances with a preposition that have been missed by the algorithm. The really disturbing figure is the one for direct objects with preposition. The algorithm has missed 106 such objects, in total 15% of all direct objects, plus 8 direct objects without a preposition. Most of the latter errors are instances where the algorithm mixes up subject and direct object (cf 17b above). The number of such errors is however low, considering the fact that Swedish often has OSV order for reasons of textual coherence.

5. Some analysis errors occuring in the material

However, the most interesting results of the application of the algorithm are perhaps the errors. What type of structures can this simplistic method not manage? Can it be improved and developed? Would more elaborate systems like those mentioned in section 1 do better? What kind of information - syntactic, semantic, pragmatic - would be needed in a lexically based parser? I will here give one example of every error type occuring in the material and discuss what kind of extra information would 'rescue' them.

The examples below are all chosen because they are typical for the kind of errors they represent. As can be seen, many sentences are quite weird and some are even ungrammatical, but that is not an argument for excluding them from the corpus. The aim of my algorithm is to handle unrestricted text; the strange sentences have occured in normal texts and should thus be given an analysis. (Cf section 6.) As a matter of fact, there are also ungrammatical sentences that get correct functional assignment by the algorithm.

Below each example I give the correct analysis and an English translation. As a comparison I then show a structurally analogous sentence, where the algorithm's analysis would be the correct one. It is clear that in most cases the latter sentence represents a more common pattern.

Erroneous subject:

(19) Plötsligt, mitt på den sterila slätten, buktar gräsmattor
(Adverbial) FV1 Subj of FV1
och står en byggnad som ser ut som ett rymdskepp.
Conj FV2 (Subj of FV2)

'Suddenly, on the sterile plain, bend lawns
and stands a building that looks like a space ship.'

Analyzed in analogy with:

(20) På slätten buktar gräsmattor och pryder sin omgivning.

'On the plain bend lawns and adorn their surroundings.'

The algorithm says that if a finite verb comes immediately after a sentence conjunction, the subject of the preceding clause is the subject of that verb. (There is no number agreement between subject and verb in Swedish.) The error in 19 would be solved by the extra information that 'står' (stands) is intransitive, so that 'en byggnad...' cannot be a direct object, and, for semantic reasons (however they are to be specified), cannot be a temporal or locative adverbial; thus it must be the subject. At the same time I find it hard to believe that the selectional restrictions of a carefully designed lexicon would allow lawns to be the subject of the verb bend.

Mixing of subject and direct object:

(21) Större betydelse än riksdagen har under hösten
(Object?) FV (Adverbial)
två andra processer.
(Subject)

'Greater importance than the parliament have during autumn two other processes.'

Analyzed in analogy with:

(22) Större universitet än Linköping har under hösten två andra problem.

'Larger universities than L. have during autumn two other problems.'

Sentence 21 has a verb phrase that is as problematic to a human as to a computer. Is the main verb a simply 'har' (has) with a nominal direct object, or is it 'har ... betydelse (än)' (is of ... importance (than))? Most analyzers would probably prefer the second solution, which would have to be given as an entity in a lexicon, with slot for comparative adjective and all. But the situation is then messed up further by the splitting and fronting of the last part of the phrase. It is highly unlikely than any lexical look-up mechanism would be able to restore this discontinuous constituent.

Erroneus direct object:

(23) Jag glömmer aldrig när Palme var i Frankrike.
Subj FV Advl (Object)

'I never forget when Palme was in France.'

Analyzed in analogy with:

(24) Jag glömmer aldrig när jag skriver upp saker i almanackan.

'I never forget when I write things in my calendar.'

On rare occasions, a temporal clause can appear as a direct object of some verbs. Those verbs can however also take temporal clauses as adverbials. I can see no generalizable way of telling when the when-clause is an adverbial and when it is an object in sentences like 23 and 24. The adverbial reading must have an overwhelming probability in its favour.

Erroneous direct object with complex verb:

(25) ... när den ena fick tag i ett snöre.
(Subj) (FV) (Object)

'... when one got hold of a string.'

Analyzed in analogy with:

(26) ... när den ena fick buggar i ett program.

'... when one got bugs in a program.'

The main verb is the whole phrase 'fick tag i' (got hold of) with 'i' as a verbal particle, but it is interpreted with the noun 'tag' as direct object and 'i' as a preposition starting a prepositional phrase incorporating what should really be the direct object. All those complex verbs - and there are many of them - must be listed as idioms, but there remains the problem that some of them can have literal readings as well.

Erroneous indirect object:

No instances in the material.

Missed prepositional direct object:

(27) ... det avvek från vad parterna träffat förlikning om, ...
Subj FV (Object)

'... it deviated from what the parties had settled, ...'

Analyzed in analogy with:

(28) Han avvek från anstalten.

'He deviated from the prison.' = 'He escaped.'

In its concrete meaning, a verb like 'avvek' is constructed with a locative adverbial, but it can also have a transferred meaning, where the locative reading of the complement seems less natural and an object interpretation is closer at hand. In these cases, there are mostly clear instances at either end and a grey zone in the middle.

Missed prepositional indirect object:

(29) (Det vet man väl) vad han gör med oss.
Obj Subj FV (Indir Obj)

'(One sure knows) what he does to us.'

Analyzed in analogy with:

(30) Det vet man väl vad han gör med kniven.

'One sure knows what he does with the knife.'

The particle verb 'gör med' with a sense of doing something against somebody/something has a particle identical to the instrumental preposition. A lexicon might tell us that knives are typical instruments and persons are not, and thus guide the analyses of 29 and 30; but how would it handle a sentence like 'One sure knows what he does with them'? A full anaphora resolution is necessary to decide whether the antecedent of 'them' is animate or not and, consequently, whether 'them' is to be analyzed as indirect object or instrumental adverbial. This does not always help either: 'books' are certainly not instrument in 'What will you do with all your books when you move?'

Many of the errors (21, 25, 27, 29) exemplify a general tendency. Verbs are often used with a transferred meaning that almost always implies an abstraction in comparison to their 'prototypical' meaning, or the meaning of the predicate is given by an 'empty' verb (be, do, have, get) plus a complement. In both cases, the pattern of functional roles of the verb/predicate is often affected.

Thus, from all this we can conclude that most of the functional assignment can be done without any recourse to lexicon at all, and for those cases that remain, the information necessary for a correct assignment is of a complex and often rather dubious nature. From the error analysis and from the senteces with 'read' in ex. 6, we can see that it is an enormous task to think of all the possible constructions, to find minimal properties that are not too ad hoc and that can allow the desired sentences and exclude the others, and to build a system that can check all these possibilities and keep track of what it is doing. To get both readings of 6c for instance, we would have to build into a parser the knowledge of the world that (in this culture) buses, but not trains or airplanes, can be used as moving signboards for advertising, and that it is thus possible either to sit on a bus and read, or to stand beside it and read on it. We must ask ourselves if this is the kind of information we would want to have in a syntactic parser and we must be aware of the fact that there will always be examples that we cannot manage. We must also acknowledge the fact that different purposes have different demands on the performance of parsers.

6. Parsers adjusted to different purposes

Parsing is no longer only a research enterprise in itself, but will more and more become a necessary prerequisite for other kinds of theoretical research or practical applications, and that fact must influence the choice of parsing algorithm. The important question of what to do with ungrammatical input should also depend on the purpose of the parser.

A parser that is used to test a specific linguistic theory should reject all input that is not in accordance with what the theory predicts. A broad coverage of language is not seen as very

important and there is a clear distinction between sentences that the parser can and can not handle, a distinction which not always corresponds to the intuitions of a human language user. Ungrammatical sentences are rejected, mostly without indication of what caused their ungrammaticality.

A parser that is part of a grammar check or critiquing system of some editor should note all deviant sentences and perhaps even suggest how they should be corrected. It can also note stylistic features and compute frequencies for single words as well as constructions. Such parsers are supposed to cover large parts of the language and to find and diagnose literally all instances of ungrammaticality. This makes their task considerably more difficult than that of the model testers, and truly there are no fullfledged critiquing systems available at present even if some attempts seem promising.

A parser that is to be used in connection with e g a question-answering system must have the ability to decide not only what is grammatical but also what makes sense. It will need more semantics than the other parsers, often in the form of knowledge about the database to which it is connected. Questions of grammaticality can be of less importance as long as the input is interpretable.

The work with large-scale corpora that is emerging in today's linguistics demands of a parser that it can analyze sentences, add structural information, and always give a best-possible analysis, with or without at the same time signalling if there is something wrong with the sentence. Its input is more or less unrestricted text and its output can be used for many forms of linguistic research, also of the model-testing kind, as well as for building and updating databases etc. In this connection there are many purposes for which a fast, simple, and robust parser of the kind suggested here is precisely what is needed.

References:

Brodda, Benny, 1987. Tracing Turns in the London-Lund Corpus with BetaText. In: Proceedings from the ALLC conference, Goteborg June 1987. Forthcoming, Sprakdata, Goteborg University.

Garside, Roger, 1987. The CLAWS wordtagging system. In: Garside - Leech - Sampson, The Computational Analysis of English. London: Longman.

Green, B.B. - Rubin, G.M., 1971. Automatic grammatical tagging of English. Providence, R.I.: Department of Linguistics, Brown University.

Jensen, Karen - Binot, Jean-Louis, 1986. Disambiguating prepositional phrase attachments by using on-line dictionary definitions. IBM Research Report RC 12148, Yorktown Heights.

Källgren, Gunnel, 1984a. HP - A Heuristic Finite State Parser Based on Morphology. In: Sågvall-Hein, Anna (ed.) De nordiska datalingvistikdagarna 1983, p. 155-162. Uppsala universitet 1984.

Källgren, Gunnel. 1984b. Automatisk excerpering av substantiv ur löpande text. Ett möjligt hjälpmedel vid automatisk indexering? IRI-rapport 1984:1. Institutet för Rättsinformatik, Stockholms universitet.

Källgren, Gunnel, 1985. A Pattern Matching Parser. In: Togeby, O. Papers from the Eight Scandinavian Conference of Linguistics. Copenhagen University.

Källgren, Gunnel. 1986. The Role of Lexicon in Computerized Analysis of Natural Language. In Dahl (ed.) Papers from the Ninth Scandinavian Conference of Linguistics. Stockholm University.

Karlsson, Fred, 1985 (ed.). Computational Morphosyntax. Report on Research 1981-1984. University of Helsinki, Publications No. 13 1985. Helsinki.