INTERPRETING SYNTACTICALLY ILL-FORMED SENTENCES

Leonardo LESMO and Pietro TORASSO

Dipartimento di Informatica - Universita' di Torino
Corso Massimo D'Azeglio 42 - 10125 Torino - ITALY

ABSTRACT

The paper discusses three different kinds of syntactic ill-formedness: ellipsis, conjunctions, and actual syntactic errors. It is shown how a new grammatical formalism, based on a two-level representation of the syntactic knowledge is used to cope with ill-formed sentences. The basic control structure of the parser is briefly sketched; the paper shows that it can be applied without any substantial change both to correct and to ill-formed sentences. This is achieved by introducing a mechanism for the hypothesization of syntactic structures, which is largely independent of the rules defining the well-formedness. On the contrary, the second level of syntactic knowledge embodies those rules and is used to validate the hypotheses emitted by the first level. Alternative hypotheses are obtained, when needed, by means of local reorganizations of the parse tree. Sentence fragments are handled by the same mechanism, but in this case the second level rules are used to detect the absence of one (or more) constituents.

INTRODUCTION

In the last years we have been involved in building a natural language (Italian) interface toward a relational database. Even if this research required to consider issues relative to knowledge representation (Lesmo et al 83) and query optimization (Lesmo et al, in press), our main concern was to devise efficient parsing techniques (Lesmo et al 81, Lesmo & Torasso 83).

The term "efficient", when applied to language processing, can take a number of different meanings, ranging from pure processing speed to the ability to analyze fragments of text, to the flexibility that characterizes the behavior of the parser. We believe that all facets of efficiency are worth being pursued, but if the communication between the man and the machine has to occur in a really natural fashion, the robustness of the parser, i.e. its ability to cope with unforeseen inputs must receive the greatest attention. It is important to realize that "unforeseen" is assumed her to refer to the syntactic form of the input sentence: of course, also inputs that are unexpected from a semantic point of view should be handled properly, but, since usually the syntactic knowledge acts as a filter between the reception of the input and the subsequent stages of the analysis, the first problem that must be faced is the following: how can the parser be prevented from rejecting sentences that are syntactically ill-formed, but could be interpreted correctly if they are passed to the other components of the system?

Alternatively, the problem can be stated as: how to foresee every interpretable input? Marcus (1982) envisages the following alternatives:

a) the use of special "un-grammatical" rules, which explicitly encode facts about non-standard usage
b) the use of "meta-rules" to relax the constraints imposed by classes of rules of the grammar
c) allowing flexible interaction between syntax and semantics, so that semantics can directly analyze substrings of syntactic fragments or individual words when full syntactic analysis fails.

Even if we agree in stating the importance of a strong interaction between syntax and semantics, our approach is quite different from c) (as well as from the other ones). For this reason, and in spite of the fact that a detailed description of the parser's operating principles has been given elsewhere (Lesmo & Torasso 83), the next section is devoted to an introduction to the basic ideas that led to the design of the syntactic knowledge source. The subsequent sections will cover some phenomena which are related with ill-formedness of sentences, namely: ellipsis, conjunctions, and some types of actual syntactic errors.

GRAMMARS AND NATURAL LANGUAGE

It is widely accepted (see Charniak 81) that syntactic knowledge consitutes one of the foundations needed to build natural language interpreters. Various kinds of grammatical formalisms have been devised to represent in efficient, flexible and perspicuous way the syntactic knowledge (Winograd 83). Even if the formalisms are quite different, the main characteristic shared by all grammars is that they are prescriptive (or normative) in nature. A grammar defines what a sentence is, that is it spec

what sequences of words are acceptable. This is in sharp contrast with the normal use of language, which has, as its main purpose, the communication of something. Of course all grammars can be (and have been) augmented in order to build a representation of the meaning of the sentences (i.e. something that should be able to carry most of its communicative contents), but a meaning can only be obtained for correct sentences.

Some efforts have recently been devoted to extending the coverage of grammars, in order to deal also with ill-formed sentences (Kwasny & Sondheimer 81, Weischedel & Sondheimer 82, Granger 82). This is usually done by relaxing the constraints imposed by some rules of the grammar, by adding new rules to take care of some kinds of ill-formedness, or by allowing the semantics to intervene when the syntax is not able to process the input. However, most of these approaches present some problems: either the perspicuousness and the readibility of the grammar is reduced or the control structure of the analyser is made considerably more complex.

The sources of ill-formedness can be grouped in three classes: ellipsis, conjunctions, and syntactic errors.

In the case of ellipsis, a fragment such as "John" or "probably" can be understood by a human listener without any particular difficulty, provided that a particular context is given. On the other hand, it is apparent that those fragments are not consistent with the rules defining the well-formed sentences.

Similar problems arise in case the grammar attempts to cope with conjunctions. In general, ellipsis is meaningful just in case a context external to the expression to analyse is assumed to exist. The situation with conjunctions is rather different: in some sense, the context that must be used to interpret a conjunct is given by the previous conjunct(s), so that it is expressed inside the sentence that has to be analysed. The difficulty in the analysis of conjunctions depends on the fact that not only the second conjunct is often ill-formed (if it is considered as a standing-alone sentence), but it is the particular form of ill-formedness that provides the analyzer with the piece of information needed to decide what is the syntactic role of that conjunct (or, if we assume that the result of the syntactic analysis is represented in form of a tree, to decide where the constituent expressed by the conjunct has to be appended in the syntactic tree). For this reason, in the following sentences the second conjuncts have quite different roles:

John loves Mary and Susy (1)
John loves Mary and Susy Fred (2)
John loves Mary and hates Violet (3)

Thus, as in the case of ellipsis, a syntactic analyser designed to handle conjunctions must be able to operate on ill-formed fragments, but with the additional difficulty of modifying the parse tree on the basis of the type of ill-formedness.

The last source of ill-formedness that we will consider are the syntactic errors. Differently from the previous cases, it is almost impossible to list all possible mistakes that a person could make in writing a sentence. Probably, most of them can not be considered as syntactic errors (e.g. misspelling of words or wrong markers for a given case of a verb), but there are also errors that have purely syntactic grounds. Some noticeable examples are agreement errors, ordering errors and errors in verb tenses. An examples of each of them is reported below:

John love Mary (4)
John is going probably to home (5)
Yesterday I have eaten a good cake (6)

Even if a more detailed discussion appears in the fifth section of this paper, it is worth noting here three points:

- most native English speakers will probably never make such errors, but, firstly, they could easily be made by non-native speakers and, secondly, at least the error exemplified in (4) could result from a typing error
- errors of that kind are more frequent in Italian, since it is richly inflectional
- even if the first and third type of errors can be (more or less) easily handled by means of relaxation techniques (Kwasny & Sondheimer 81), this is not the case for ordering errors; this is due to the fact that the agreement and tense constraints are expressed "explicitly" in the grammar (e.g. by an augmentation), whereas the order is specified implicitly (i.e. rigidly embodied in the grammar itself).

The analysis of the problems mentioned in this section, together with some other considerations that are not worth being discussed extensively here (regarding, for instance, garden paths) led us to the design of a formalism for representing the syntactic knowledge that splits it into two levels. The first level contains a set of rules that, in our intention, characterize the meaningful sentences. It can be questioned whether rules regarding meaning can be considered as syntactic rules. Our opinion is that the syntactic categories associated with natural language words have a strong semantic bias (see, for a thorough discussion of this thesis (Lyons 77, Chapt.11)). For this reason, we defined a set of node types that have to be used in building the tree representing the syntactic structure of the sentence. These node types (reported in table 1) are associated with the syntactic categories and the topological constraints that gov

REL	Relation	Verbs, copulas
REF	Referent	Nouns, pronouns
CONN	Connector	Prepositions, conjunctions
DET	Determiner	Articles demonstrative adjectives, adjectival question words
MOD	Adverbial Modifier	Adverbs
ADJ	Adjectival Modifier	Adjectives

Table 1 - The node types: The first column contains the name (actual and extended); the second one contains the classical syntactic categories associated with the node type

ern the attachment of nodes constitute the basic filter which selects the "meaningful" fragments of sentence. As an example of this kind of constraints, it is unreasonable to assume that an ADJ node can be attached elsewhere than a REF node (with the exception of verbs having a copulative function, e.g. to be, to seem, to taste etc.). For this reason, independently of its position in the sentence, we can exclude some kinds of constructs (e.g. ADJ-ADJ attachment) as meaningless.[1] When a rule of the first set is executed it (normally) involves the creation of a new node (possibly more than one) and its attachment to the syntactic tree which was built up to that time.

Because of the limited knowledge used to hypothesize the attachment point, it can often happen that the parser made the wrong choice. Such an error can be detected by using two different knowledge sources: higher-level syntactic constraints and semantics. The first of them contains the rules that define the well-formedness of sentences (in particular gender-number agreements rules and ordering rules) whereas the second knowledge source tells whether an attachment is semantically acceptable (of course, even if a REF-ADJ attachment is consistent with the topological constraints, not all adjectives can be used to qualify a given noun). The semantic checks are done accessing a semantic net organized in two levels: the first of them (external) concerns the acceptable surface structures (e.g. case frames for verbs), whilst the second one (internal) is concerned with the actual semantics of the domain (e.g. subsetting among classes).

[1] it must be noted that the rules embodying these constraints are expressed in procedural form.Even if the lack of a declarative representation makes more difficult the design and the maintenance of the rules, they are made more efficient in terms of execution time by taking into account the context where the word occurs (involving a limited one word lookahead).

Because of the frequency of this kind of wrong hypothesization, an effective computational tool must be used to restructure the tree: this tool consists in what we called "natural changes", which are simple pattern-action rules able to move around constituents; their purpose is to provide the parser with an alternative hypothesis when a given one has failed. Whereas the natural changes are triggered the same way both in case the inconsistency is syntactic and semantic, different courses of action take place if the changes cannot produce any acceptable alternative hypothesis: if the error is of syntactic type than the first hypothesis is maintained but a warning message is sent to the user; if the error is semantic, then the current interpretation of the fragment is considered unacceptable and, in case one or more choice points were previously met, the parser backtracks, otherwise the analysis fails. More details about the use of backup, as well as about other topics related with the parsing strategy, can be found in (Lesmo & Torasso 83).

A problem which must be faced when a natural change is stimulated is the choice of the best interpretation. Let us suppose that an agreement between an adjective and a noun is violated. In this case the natural change MOVE UP tries to attach the adjective to a REF node which is at a higher level with respect to the REF which the adjective is currently attached to. The new attachment stimulates the rules of the second set (that is the rules verifying the agreement and the word ordering) and the semantic ones. It is possible that the semantic rules signal that the new attachment is not admissible from a semantic point of view. At this point, if no alternative attachment is possible, the system has to consider the first interpretation as the best one since it violates only the "weak" syntactic constraints.

ELLIPSIS

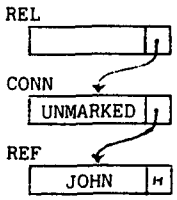
"Ellipsis" is a greek word (elleipsis) roughly corresponding to "lack, omission", that is used, to take a dictionary definition, to stand for "omission of one or more words that can easily be subsumed". Even if all components of the definition are fundamental, we want to stress the presence of the adverb "easily". It is consistent with the observation that, whereas other phenomena occurring in natural language (e.g. garden path) require a conscious effort in the listener, elliptical sentences are understood without any difficulty. On the other hand, most current grammatical formalisms are not able to account for this ease in understanding ellipsis; it must be noted the importance that is often laid on the ability to decide as soon as possible what is the allowable form of a given constituent (Buchenko et al. 83). This is due to the necessityof triggering in advance a suitable re-

stricted set of grammar rules. In our case this is not required: the first-level rules will work the same way independently of the global context where a given word or constituent occurs (this is not true for "local" contexts in the current version of the system: see note 1); the consistency with the rules which govern the construction of well-formed sentences will be tested afterwards. This is particularly useful for handling elliptical fragments. Let's see through a pair of examples what is the behaviour of the parser in such sistuations.

Example (1) is reported below:

John (1)

The rules associated with the category "noun" (note that the first-level rules are grouped in packets associated with syntactic categories), in case the analysis is at the beginning of the sentence, cause the building of the sentence reported below:



When the end of the sentence in encountered, the structure is recognized as being incomplete and a pattern matching procedure applied to any preceding question can reconstruct its actual meaning. What must be noticed is that the first-level syntactic rules used to analyze the fragment are exactly the same that are used to analyze complete and correct sentences.

CONJUNCTIONS

The kind of processing that occurs in handling conjunctions requires the introduction of rather different constraints. The first interpretation produced for sentences 3) and 4) after the fragment "John loves Mary and Susy" has been analyzed is reported in fig. 1a. This interpretation is confirmed when the end of sentence 3) is encountered (so that the final structure is the one shown in fig. 1a). On the contrary, when the name "Fred" is scanned in sentence 4), it cannot be attached to "Susy" (excluding the possibility that "Fred" is her family name) and the attempt to move it up to "loves" causes a semantic error (three unmarked case for "love"). At this point another "natural change" is triggered, which handles conjunctions. It tries to move up the "and" node, producing the structure of fig.1b which is accepted as the correct one. Note, however, that this kind of natural change is much more complex than the standard ones. For example, in the reported examples two new nodes have to be built: the empty REL node (this is done easily since only two nodes of the same type can be connected via "and")

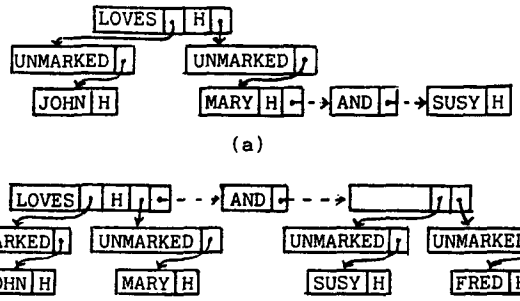


Fig.1 - The parse trees for sentence 3) (fig.1a) and sentence 4 (fig.1b).

and the "UNMARKED" connection (for which an explicit request of creation and attachment must be issued).

A final observation regards the fact that the parser assumes that the first acceptable interpretation is the right one. This implies that a sentence of the form (see EX4 in Huang 83, pag.82) "The man with the telescope and the woman with the umbrella kicked the ball" would be interpreted as "The man with the telescope and with the woman with the umbrella kicked the ball", that is not the most natural interpretation for a human listener. However, Italian always expresses explicitly the number of the verb (i.e. plural in this case), so that the Italian translation of the sentence would be analyzed correctly.

SYNTACTIC ERRORS

The system tolerates and possibly recovers the following different kinds of errors:

- lexical errors
- agreement errors
- errors in the ordering of the constituents
- extra cases

(note that only the second and the third kind of errors are actual syntactic errors).

As regards the errors at the lexical level, they are detected when the morphological analyzer tries to decompose a given word in "root + suffix" form. When no decomposition is possible or none of the obtained roots occurs in the dictionary, the system asks the user about the possibility that the input word is mispelled. In the affirmative case, the user can retype the word, whereas in the opposite case the system asks the user to provide it with some pieces of information such as the syntactic category of the word, its normalized form (i.e. its root), the gender, the number, etc.; moreover the system asks what semantic object the word refers to. In this way the analysis of the sentence can go on and possibly an interpretation is constructed. However, it has to be pointed out that the information provided by the user during the

analysis of the sentence is not always sufficient for the system to complete the analysis. In fact, the current version of the system has not the capability of restructuring the semantic net dynamically, so that the system can continue the analysis only when the semantic object denoted by the unknown word is already present in the net.

As regards "agreement errors" there is a large variety of error types grouped under this label:

a) a first kind refers to the agreement in number and gender between the noun and the determiner and between the noun and the adjectives. It is worth noticing that such kind of errors is uncommon in Italian, because the suffixes for male and female and for singular and plural are in many cases quite different.
b) A slightly more frequent error concerns the agreement in number, gender and person between the subject and the verb. Since in Italian the suffixes indicating the different persons of the verb, its tense and mood are quite different, people whose mother tongue is Italian usually do not make this kind of mistake.
c) Another kind of agreement refers to the relationships existing between the moods and the tenses of the verbs occurring in the main sentences and its subordinates. The rules, which are quite complex since they derive from the "consecutio temporum" of Latin, are often violated so that this kind of error must be tolerate by the system. In this case the procedure which has the task of verifying the agreement emits a warning message when the rules are violated, but, contrarily to cases a) and b), it does not try to restructure the parse tree via "natural changes", since in most cases no alternative interpretation exists.

The framework we have provided is particularly useful for treating errors in the ordering of the constituents, in fact the order is checked only when a given sentence (possibly a subordinate) has been completed. This happens when the REL node that heads the clause (main or subordinate) is closed, that is a punctuation mark is encountered or a new node is attached to a node which is (in the parse tree) at a level higher than the REL currently analized. Before stimulating the ordering rules, the system checks that the case frame of REL has been correctly filled, that is all the cases attached to REL are compatible with the head and among them. Just in this case a set of rules is activated depending on the sentence type (it is apparent that the constituent order is different in a declarative, interrogative or relative clause). Each rule represents a legitimate ordering of the constituents and the rules are ordered in decreasing degree of acceptability. The rules are matched in turn against the actual case frame of the verb acting as head of the clause under examination; in case no rule matches, a warning is issued to signal the user that something has gone wrong in the ordering; anyway the interpretation of the clause obtained by accessing the semantic net is maintained and the analysis goes on if the entire sentence has not yet been scanned. A similar (but simpler) processing occurs for a REF node with respect to the adjectives attached to it.

There are also cases which are more difficult to treat than the ones involving violations in the word ordering. In fact, a sentence like "Il giornale lo ha comprato Giovanni stamattina" (literally "The newspaper it has bought John this morning") involves not only word order violations (the syntactic object occurs in the first position in the sentence), but also there is a case denoted by "lo" ("it") which duplicates the object. Such sentences are clearly incorrect from a syntactic point of view as well as, in principle, from a semantic one (wrong case frame), but they are perfectly understandable and quite frequent because they allow one to identify as focus of the utterance the object without passivizing the sentence.

The treatment of such kinds of errors requires only relatively inexpensive modifications to the way the semantic net is accessed. It is worth noticing, in fact, that the syntactic object ("il giornale") is attached to a REL node which is empty when this attachment is performed. The semantic and agreement check procedures are stimulated but are immediately suspended since the REL node is empty. Similarly the pronoun "lo" is attached to the REL and the corresponding check procedures are suspended. When the REL node has been filled with "comprato" the suspended checks are resumed. The semantic procedure is able, by inspecting the semantic net, to state that "giornale" may fill the "object" role so that when the previously suspended semantic check is executed, it concludes that "lo" ("it") cannot be attached to the REL filled with "comprare" ("buy") since the object role has already been filled.

Instead of rejecting the current interpretation by stimulating the natural changes and possibly the backup mechanism, a modification of the parsing strategy consists in attaching a warning to the REF node containing the pronoun "lo" and in going on with the sentence analysis. When the sentence has been completely scanned and, consequently, it is possible to perform a global check on the actual case frame of "comprare", the semantic procedure decides that "lo" is simply a repetition of the object and therefore it may be disregarded. In this way the interpretation of the sentence is possible, but the warning attached to the REF node containing "lo" is output to the user.

CONCLUSIONS

The paper presents a parsing strategy able to cope with different kinds of syntactic ill-formedness: ellipsis, conjunctions, syntactic errors. Some examples are reported to show that the adopted formalism allows the parser to analyse ill-formed fragments without substantial changes to the rules used to analyse correct sentences.

However, some problems still deserve further attention. First of all, in case of ill-formed sentences it is often possible to assign more than one interpretation to the sentence (e.g. in "The boy love the girl" the subject can be considered plural - missing "s" in "boy" - or singular - missing "s" in "love"); this can also happen for correct sentences (see the last example in the section on CONJUNCTIONS). The current version of the system should be enhanced both by taking into account contextual information (which could be useful in the first case) and by weighing in some way the output of the semantic component (which, today, is categorical: yes or no).

As regards the context, the experiments we made on the parser refer to isolated sentences, so that the "pattern matching" procedure we referred to in the section on ELLIPSIS (see the example "John") is neither implemented nor designed. Our belief is that the two components (pattern matcher and parser) are quite independent each other, but we are planning to address also issues connected with discourse analysis.

Last but not least, some problems are more strictly connected with the basic parser design. Some English sentences break a locality principle embodied in the first-level syntactic rules. An example is given by "What architect do you know who likes the balalaika" (see Winograd 83, pag.136). We are currently studying this problem, whose solution will involve a change in the final representation as well as in the rule packets.

The current version of the parser, that runs on a VAX-11/780 under the UNIX operating system and is implemented in FRANZ LISP, includes the mechanisms for detecting and recovering the lexical, agreement, and word ordering errors, whereas the "extra cases", in the sense explained above, are currently being implemented.

REFERENCES

Bachenko J., Hindle D., Fitzpatrick: Constraining a Deterministic Parser. Proc. AAAI-83 (1983)8-11.

Charniak E.:Six Topics in Search of a Parser: An Overview of AI Language Research. Proc. 7th IJCAI Vancouver B.C. (1981), 1074-1087.

Huang X.: Dealing with Conjunctions in a Machine Translation Environment. Proc. 1st Conf. ACL-Europe, Pisa (1983), 81-85.

Granger R.H.: Scruffy Text Understanding: Design and Implementation of "Tolerant" Understanders. Proc. 20th ACL, Toronto (1982), 157-160.

Kwasny S.C., Sondheimer N.K.: Relaxation Techniques for Parsing Grammatically Ill-Formed Input in Natural Language Understanding Systems. AJCL 7 (1981), 99-108.

Lesmo L., Magnani D., Torasso P.: A Deterministic Analyzer for the Interpretation of Natural Language Commands. Proc. 7th IJCAI, Vancouver B.C. (1981), 440-442.

Lesmo L., Siklossy L., Torasso P.: A Two-Level Net for Integrating Selectional Restrictions and Semantic Knowledge. Proc. IEEE Int. Conf. on System, Man and Cybernetics, India (1983), 14-18.

Lesmo L., Torasso P.: A Flexible Natural Language Parser based on a Two-Level Representation of Syntax, Proc. 1st Conf. ACL-Europe, Pisa (1983), 114-121.

Lesmo L., Siklossy L;, Torasso P.: Semantic and Pragmatic Processing in FIDO: A Flexible Interface for Database Operations. Accepted for Publication on Information Systems.

Lyons J.: Semantics. Cambridge Univ. Press (1977).

Marcus M.: Building Non-Normative Systems: The Search for Robustness: An Overview. Proc. 20th ACL, Toronto (1982), 152.

Weischedel R.M., Sondheimer N.K.: An Improved Heuristic for Ellipsis Processing. Proc. 20th ACL, Toronto (1982), 85-88.

Winograd T.: Language as a Cognitive Process; Vol.1 Syntax. Addison Wesley (1983).