# INCREMENTAL DEPENDENCY PARSING

Vincenzo Lombardo
*Dipartimento di Informatica - Universita' di Torino*
C.so Svizzera 185 - 10149 Torino - Italy
e-mail: vincenzo@di.unito.it

## Abstract

The paper introduces a dependency-based grammar and the associated parser and focusses on the problem of determinism in parsing and recovery from errors. First, it is shown how dependency-based parsing can be afforded, by taking into account the suggestions coming from other approaches, and the preference criteria for parsing are briefly addressed. Second, the issues of the interconnection between the syntactic analysis and the semantic interpretation in incremental processing are discussed and the adoption of a TMS for the recovery of the processing errors is suggested.

## THE BASIC PARSING ALGORITHM

The parser has been devised for a system that works on the Italian language. The structure that results from the parsing process is a dependency tree, that exhibits syntactic and semantic information.

**The dependency structure:** The structure combines the traditional view of dependency syntax with the feature terms of the unification based formalisms (Shieber 86): single attributes (like number or tense) appear inside the nodes of the tree, while complex attributes (like grammatical relations) are realized as relations between nodes. The choice of a dependency structure, which is very suitable for free word order languages (Sgall et al. 86), reflects the intuitive idea of a language with few constraints on the order of legal constructions. Actually, the flexibility of a partially configurational language like Italian (that can be considered at an intermediate level between the totally configurational languages like English and the totally inflected free-ordered Slavonic languages) can be accounted for with a relaxation of the strong constraints posed by a constituency grammar (Stock 1989) or by constraining to a certain level a dependency grammar. Cases of topicalization, like

**un dolce di frutta ha ordinato il maestro**
*a cake with fruits has ordered the teacher*

and in general all the five permutations of the "basic" (i.e. more likely) SVO structure of the sentence are so common in Italian, that it seems much more economical to express the syntactic knowledge in terms of dependency relations.

Every node in the structure is associated with a word in the sentence, in such a way that the relation between two nodes at any level is of a head&modifier type. The whole sentence has a head, namely the verb, and its roles (the subj is included) are its modifiers. Every modifier in turn has a head (a noun, which can be a proper, common or pro-noun, for participants not marked by a preposition, a preposition, or a verb, in case of subordinate sentences not preceded by a conjunction) and further modifiers.

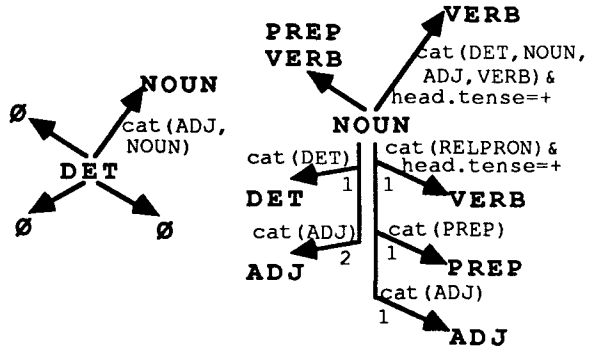
Hence the dependency tree gives an immediate representation of the thematic structure of the sentence, thus being very suitable for the semantic interpretation. Such a structure also allows the application of the rules, based on grammatical relations, that govern complex syntactic phenomena, as revealed by the extensive work on Relational Grammar.

The dependency grammar is expressed declaratively via two tables, that represent the relations of immediate dominance and linear order for pairs of categories. The constraints on the order between a head and one of its modifiers and between two modifiers of the same head are reflected by the nodes in the dependency structure. The formation of the complex structure that is associated with the nodes is accomplished by means of unification: the basic terms are originated by the lexicon and associated with the nodes. There exist principles that govern the propagation of the features in the dependency tree expressed as analogous conventions to GPSG ones.

**The incremental parser:** In the system, the semantic, as well as the contextual and the anaphoric binding analysis, is interleaved with the syntactic parsing. The analysis is incremental, in the sense that it is carried out in a piecemeal strategy, by taking care of partial results too.

In order to accomplish the incremental parsing and to build a dependency representation of the sentence, the linguistic knowledge of the two tables is

compiled into more suitable data structures, called diamonds. Diamonds represent a redundant version of the linguistic knowledge of the tables: their graphical representation (see the figure) gives an immediate idea of how to employ them in an incremental parsing with a dependency grammar.




The center of the diamond is instantiated as a node of the category indicated during the course of the analysis. The lower half of the diamond represents the categories that can be seen as modifiers of the center category. In particular, the categories on the left will precede the head, while the categories on the right will follow it (the number on the edges totally order the modifiers on the same side of the head). The upper half of the diamond represents the possible heads of the center: the categories on the right will follow it, while the categories on the left, that precede it, indicate the type of node that will become active when the current center has no more modifiers in the sentence.

The (incremental) parsing algorithm is straightforward: if the current node is of category X, the correspondent diamond (which has X as the center) individuates the possible alternatives in the parsing. The next input word can be one of its possible modifiers that follow it (right-low branch), its head (right-up branch), another modifier of its head, i.e. a sister (right-up branch and the following left-down one in the diamond activated immediately next), or a modifier of its head's head, an aunt (left-up branch).

The edges are augmented with conditions on the input word (*cat* is a predicate which tests its category as belonging to a set of categories allowed to be the left-corner of the subtree headed by a node of the category that stands at the end of the edge). Constraints on features are tested on the node itself or stored for a subsequent verification.

Which edge to follow in the currently active diamond is almost always a matter of a non deterministic choice. Non determinism can be handled via the interaction of many knowledge sources that use the dependency tree as a shared information structure, that represents the actual state of the parsing. Such a structure does not contain only syntactic, but also semantic information. For example, every node associated with a non functional word points to a concept in a terminological knowledge base and the thematic structure of the verb is explicitly represented by the edges of the dependency tree.

## PARSING PREFERENCES

Many preference strategies have been proposed in the literature for guiding parsers (Hobbs and Bear (1990) present a review). There are some preferences of syntactic (i.e. structural) nature, like the Right Association and the Minimal Attachment, that were among the first to be devised. Semantic preferences, like the assignment of thematic roles to the elements in the sentence[1] can contradict the expectations of the syntactic preferences (Schubert 1984). Contextual information (Crain, Steedman 1985) has also been demonstrated to affect the parsing of sentences in a series of psycholinguistic experiments. Lexical preferencing (Stock 1989) (van der Linden 1991) is particularly useful for the treatment of idiomatic expressions.

Parsing preferences are integrated in the framework described above, by making the syntactic parser interact with condition-action rules, that implement such preferences, at each step on the diamond structure. This technique can be classified under the weak integration strategy (Crain, Steedman 1985) at the word level. The rules for the resolution of ambiguities that belong to the various knowledge sources analyze the state of the parsing on the dependency structure and take into account the current input word. For example, in the two sentences

**a) Giorgio le diede con riluttanza una ingente somma di denaro**
*Giorgio (to) her gave with reluctance a big amount of money*
**b) Giorgio le diede con riluttanza a Pamela**
*Giorgio them gave with reluctance to Pamela*
the pronoun "le" can be a plural accusative or a singular dative case. In an incremental parser, when we arrive to "le" we are faced with an ambiguity that can be solved in a point which is arbitrarily ahead (impossibility of using Marcus' (1980) bounded

[1]As we have noted in the beginning, this is not an easy task to accomplish, since flexible languages like Italian feature a hardly predictable behavior in ordering: such assignments must sometimes be revised (see below).

lookahead), when we find which grammatical relation is needed to complete the subcategorization frame of the verb. Contextual information can help in solving such an ambiguity, by binding the pronoun to a referent, which can be singular or plural. Of course there could be more than one possible referent for the pronoun in the example above: in such a case there exist a preference choice based on the meaning of the verb and its selectional restrictions, and, in case of further ambiguity, a default choice among the possible referents. This choice must be stored as a backtracking point (in JTMS style) or as being an assumption of a context (in ATMS style), since it can reveal to be wrong in the subsequent analysis. The revision of the interpretation can be accomplished via a reason maintenance system.

## INTEGRATION WITH A REASON MAINTENANCE SYSTEM

Zernik and Brown (1988) have described a possible integration of default reasoning in natural language processing. Their use of a JTMS has been criticized because of the impossibility to evaluate the best way in presence of multiple contexts, that are available at a certain point of the parsing process. This is the reason why more recent works have focussed on ATMS techniques (Charniak, Goldman 1988) and their relations to chart parsing (Wiren 1990). ATMS allows to continue the processing, by reactivating interpretations, which have been previously discarded.

Currently, the integration with a reason maintenance system (which can possibly be more specialized for this particular task) is under study. The dependency structure contains the short term knowledge about the sentence at hand, with a "dependency" (in the TMS terminology) net that keeps the information on what relations have been inferred from what choices. Once that new elements contradict some previous conclusions, the dependency net allows to individuate the choice points that are meaningful for the current situation and to relabel, according to the IN and OUT separation, the asserted facts. In the example a) if we have disambiguated the pronoun "le" as an object, such an interpretation must be revised when we find the actual object ("a big amount of money"). One of the reasons for adopting truth maintenance techniques is that all the facts that must be withdrawn and the starting of a new analysis (in JTMS style) or to make relevant a new context in place of an old one (in ATMS) must take into account that partial analyses, not related to the changes at hand ("with reluctance" in the example), must be left unchanged. The specific substructure A, affected by the value chosen for the element B, and the element B are connected via a (direct or indirect) link in the "dependency" net. A change of value for B is propagated through the net toward all the linked substructures and, particularly, to A, which is to be revised. In the example a), once detected that "le" is an indirect object, and then that its referent must be female and singular, a new search in the focus is attempted according to this new setting. Hence, the revision process operates on both the syntactic structure, with changes of category and/or features values for the nodes involved (gender and number for "le") and of attachment points for whole substructures, and the semantic representation (from direct to indirect object relation), which has been previously built.

## ACKNOWLEDGEMENTS

I thank prof. Leonardo Lesmo for his active and precious support.

## REFERENCES

Charniak, E., Goldman, R. (1988). A Logic for Semantic Interpretation. In *Proceedings of the 26th ACL* (87-94).

Crain, S., Steedman, M. (1985). On not being led up the Garden Path: The Use of Context by the psychological Syntax Processor. In D. Dowty, L. Karttunen and A. Zwicky (eds), *Natural Language Parsing. Psychological, Computational, and Theoretical Perspectives*, Cambridge University Press, Cambridge, England (320-358).

Hobbs, J., Bear, J. (1990). Two Principles of Parse Preference. In *COLING 90* (162-167).

van der Linden, E., J. (1991). Incremental Processing and Hierarchical Lexicon. To appear.

Marcus, M. (1980). *A Theory of Syntactic Recognition for Natural Language*. MIT Press, Cambridge, Massachussets.

Schubert, L. (1984). On parsing preferences. In *COLING 84* (247-250).

Sgall, P., Haijcova, E. and Panevova, J. (1986). *The Meaning of the Sentence in its Semantic and Pragmatic Aspects*. D. Reidel Publishing Company.

Shieber, S., M. (1986). *An Introduction to Unification-Based Approach to Grammar*. CSLI Lecture Notes 4, CSLI, Stanford.

Stock, O. (1989). Parsing with flexibility, dynamic strategies and idioms in mind. In *Computational Linguistics 15* (1-19).

Wiren, M. (1990). Incremental Parsing and Reason Maintenance. In *COLING 90* (287-292).

Zernik, U., Brown, A. (1988). Default Reasoning in Natural Language Processing. In *COLING 88* (801-805).