WEDNESDAY: Parsing Flexible Word Order Languages

Oliviero Stock
Cristiano Castelfranchi
Domenico Parisi

Istituto di Psicologia
del Consiglio Nazionale delle Ricerche
Via dei Monti Tiburtini 509, 00157 Roma

ABSTRACT

A parser for "flexible" word order languages must be substantially data driven. In our view syntax has two distinct roles in this connection: (i) to give impulses for assembling cognitive representations, (ii) to structure the space of search for fillers. WEDNESDAY is an interpreter for a language describing the lexicon and operating on natural language sentences. The system operates from left to right, interpreting the various words comprising the sentence one at a time. The basic ideas of the approach are the following:

a) to introduce into the lexicon linguistic knowledge that in other systems is in a centralized module. The lexicon therefore carries not only morphological data and semantic descriptions. Also syntactic knowledge is distributed throughout it, partly of a procedural kind.

b) to build progressively a cognitive representation of the sentence in the form of a semantic network, in a global space, accessible from all levels of the analysis.

c) to introduce procedures invoked by the words themselves for syntactic memory management. Simply stated, these procedures decide on the opening, closing, and mantaining of search spaces; they use detailed constraints and take into account the active expectations.

WEDNESDAY is implemented in MAGMA-LISP and with a stress on the non-deterministic mechanism.

1. Parsing typologically diverse languages emphasizes aspects that are absent or of little importance in English. By taking these problems into account, some light may be shed on:

a) insufficiently treated psycholinguistic aspects

b) a design which is less language-dependent

c) extra- and non-grammatical aspects to be taken into consideration in designing a friendly English user interface.

The work reported here has largely involved problems with parsing Italian. One of the typical features of Italian is a lower degree of word order rigidity in sentences. For instance, "Paolo ama Maria" (Paolo loves Maria) may be rewritten without any significant difference in meaning (leaving aside questions of context and pragmatics) in any the six possible permutations: Paolo ama Maria, Paolo Maria ama, Maria ama Paolo, Maria Paolo ama, ama Paolo Maria, ama Maria Paolo. Although Subject-Verb-Object is a statistically prevalent construction, all variations in word order can occur inside a component, and they may depend on the particular words which are used.

2. In ATNSYS (Cappelli, Ferrari, Moretti, Prodanof and Stock, 1978), a previously constructed ATN based system (Woods, 1970), a special dynamic reordering mechanism was introduced in order to get sooner to a correct syntactic analysis, when parsing sentences of a coherent text (Ferrari and Stock, 1980). Besides psycholinguistic motivations, the main reason for the introduction such heuristics lay in the large number of alternative arcs that has to be introduced in networks for parsing Italian sentences.

As a matter of fact, ATN's were not originally conceived for flexible word order languages. (In the extreme free word order case, an ATN would have one single node and a large number of looping arcs, losing its meaningfulness).

Work has been done on ATN parsers for the parsing of non-grammatical or extra-grammatical sentences in English, a problem related to our one. For instance Weischedel and Black (1981) have proposed a system of information passing in the case of parsing failure. Kwasny and Sondheimer (1981) have suggested the relaxation of constraints on the arcs under certain circumstances. Nevertheless, these problems, together with that of treating idiosyncratic phenomena related to words and flexible idioms, are not easy to solve within the ATN approach.

At least two other parsers should be mentioned here.

ELI (Riesbeck and Schank, 1976) derives directly from the conceptual dependency approach. The result of the analysis is based on semantic primitives, and the analysis is governed by concept expectations. The analyzer's non-determinism is in large part eliminated by world knowledge consultation. In practice, the (scanty) syntax is considered only later, in case of difficulty.

The problem with this approach is represented by the difficulty in controlling cases of complex linguistic form.

Small's Word Expert Parser (Small, 1980) is an interesting attempt to assign an active role to the lexicon. The basic aspect of parsing, according to Small's approach, is disambiguation. Words may have large numbers of different meanings. Discrimination nets inserted in words indicate the paths to be followed in the search for the appropriate meaning. Words are defined as coroutines. The control passes from one word, whose execution is temporarily suspended, to another one and so on, with reentering in a suspended word if an event occurs that can help proceeding in the suspended word's discrimination net.

This approach too takes into little account syntactic constraints, and therefore implies serious problems while analyzing complex, multiple clause sentences.

It is interesting to note that, though our approach was strictly parsing oriented from the outset, there are in it many similarities with concepts developed independently in the Lexical-Functional Grammar linguistic theory (Kaplan & Bresnan, 1982).

3. A parser for flexible word order languages must be substantially data driven. In our view syntax has two distinct roles in this connection

- to give impulses for assembling cognitive representations (basically impulses to search for fillers for gaps or substitutions to be performed in the representations)

- to structure the space of search of fillers.

WEDNESDAY, the system presented here, is an interpreter for a language describing the lexicon and operating on natural language sentences. The system operates from left to right, interpreting the various words comprising the sentence one at a time.

The diagram for WEDNESDAY is shown in Fig.1. The basic ideas of the approach are the following:

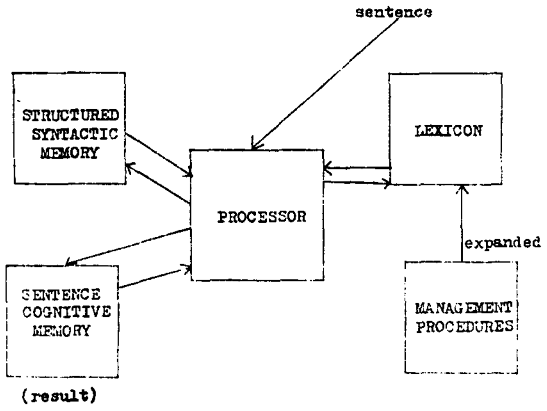

Fig.1

a) to introduce into the lexicon linguistic knowledge that in other systems is in a centralized module. The lexicon therefore carries not only morphological data and semantic descriptions. Also syntactic knowledge is distributed throughout it, partly of a procedural kind. In other words, though the system assigns a fundamental role to syntax, it does not have a separate component called "grammar". By being for a large part bound to words, syntactic knowledge makes it possible to specify the expectations that words bring along, and in what context which conditions will have to be met by candidates to satisfy them. "Impulses", as they are called in WEDNESDAY to indicate their active role, result in connecting nodes in the sentence cognitive memory. They may admit various alternative specifications, including also side-effects such as equi-np recognition, signalling a particular required word order, etc.

The advantages of this aspect of WEDNESDAY include:

- easy introduction of idiosyncratic properties of words;

- possibility of dealing with various types of non-generative forms (idioms).

b) to build progressively a cognitive representation of the sentence in the form of a semantic network, in a global space, accessible from all levels of the analysis.

A word representation forms a shred of network that is later connected with other shreds until the complete network is formed. The representation we use is neutral enough to guarantee that any changes in the format will not cause serious problems to the analyzer. In substance it can be seen as a propositional format in Polish Prefixed notation:

$$(N_x(P\ N_1\ \ldots\ N_i\ \ldots\ N_m))$$

where N_x is an instantation of predicate P, nodes $N_1\ \ldots\ N_m$ are the variables, arguments of that predicate. Some decompositional analysis is performed, leading to a possible complex set of propositions for expressing the meaning of a word.

c) to introduce procedures invoked by the words themselves for syntactic memory management. Simply stated, these procedures decide on the opening, closing, and mantaining of search spaces; they use detailed constraints and take into account the active expectations. They are, as the lexicon obviously is, dependent on the particular language; nevertheless they refer to general primitive concepts. The procedures can be looked upon as a redefinition of syntactic categories in procedural terms, based on lower level primitive functions. This can be viewed as a different perspective on this aspect of linguistics, traditionally considered in a static and taxonomic way.

To manage structured spaces in this way allows:

- to maintain a syntactic control in the analysis of complex sentence

- to keep an emphasis on the role played by the lexicon.

Fig.2 shows a space management procedure, considering two space types, S and N.

```
(*NOUN ()
      (S(COND((CANCLOSE)
                    (NON-DET(T(CLOSESPACE)
                               (*NOUN))
                          ((IS-EXPECTED N NS)
                               (OPENSPACE N))))
             ((OR(NOT(MAIN-ARRIVED))
                    (IS-EXPECTED N NS))
                          (OPENSPACE N))
             ((FAIL))))
     (N(COND((CANCLOSE)(CLOSESPACE)(*NOUN)))))
```

Fig. 2

The following memories are used by WEDNESDAY:

1) a SENTENCE COGNITIVE MEMORY in which semantic material carried by the words is continuously added and assembled. This memory can be accessed at any stage of the parsing.

2) a STRUCTURED SYNTACTIC MEMORY in which, at every computational level:

- the expectations defining the syntactic space are activated (e.g. the expectation of a verb with a certain tense for a space S)

- the expectations of fillers to be merged with the gap nodes are activated

- the nodes capable of playing the role of fillers are memorized

- there are various local and contextual indications.

4. Impulses can be of two types. A MERGE is an impulse to merge an explicitly indicated node with another node that must satisfy certain constraints, under certain conditions. MERGE is therefore the basic network assembling resource. We use to characterize the node quoted in a MERGE impulse as a "gap" node, a node that actually is merged with a gap node as a "filler" node.

A MERGE impulse can state several alternative specifications for finding a filler.

The following are specified for each alternative:

a) an alt-condit, i.e. a boolean predicate concerned with possible flag raising occurring during the process.

b) a fillertype, i.e. the syntactic characteristic of the possible filler. A fillertype can be a headlist (e.g. N), or $$MAIN, indication of the main node of the current space, or $$SUBJ, indication of the subject of the current space.

c) the indication of the values of the features that must not be in contrast with the corresponding features of the filler (i.e. an unspecified value of the feature in the filler is ok, a different value from the one specified is bad). If the value of the feature in the filler is NIL, the value specified here will be assumed.

d) a markvalue that must not be contrasted by the markvalue of the filler

e) sideffects caused by the merging of the nodes. These can be: SETFLAG, which raises a specified flag (that can subsequently alter the result of a test), REMFLAG, which removes a flag, and SUBSUBJ, which specifies the instantiation node and the ordinal number of the relative argument identifying a node. The subject of the subordinate clause (whose MAIN node will be actually filling the gap resulting from the present MERGE) will be implicitly merged into the node specified in SUBSUBJ. It should be noted that the latter may also be a gap node, in which case also after the present operation it will maintain that characteristic.

MARK is an impulse to stick a markvalue

onto a node. If the chosen node has already a markvalue, the new one will be forced in and will replace it.

MUST indicates that the current space will not be closed if the gap is not filled. Not all gaps have a MUST: in fact in the resulting network there is an indication of which nodes remain gaps.

As mentioned before, the merging of two nodes is generally an act under non-deterministic control: a non-deterministic point is established and the first attempt consists in making the proposed merging. Another attempt will consist in simply not performing that merging. A FIRST specification results in not establishing a non-deterministic point and simply merging the gap with the first acceptable filler.

By and large the internal structure of gaps may be explained as follows.

A gap has some information bound to it. More information is bound to subgaps, which are LISP atoms generated by interpreting the specification of alternatives within a MERGE impulse. When an "interesting event" occurs those subgaps are awakened which "find the event promising".

Subsequently, if one of the subgaps actually finds that a node can be merged with its "father" gap and that action is performed, the state of the memories is changed in the following way:

- in the SENTENCE COGNITIVE MEMORY the merging results in substitution of the node and of inverse pointers.

- in the STRUCTURED SYNTACTIC MEMORY the gap entity is eliminated, together with the whole set of its subgaps.

Furthermore if the filler was found in a headlist, it will be removed from there.

Note that while the action in the SENTENCE COGNITIVE MEMORY is performed immediately, the action in the STRUCTURED SYNTACTIC MEMORY may occur later.

One further significant aspect is that with the arrival of the MAIN all nodes present in headlists must be merged. If this does not happen the present attempt will abort.

5. WEDNESDAY is implemented in MAGMA-LISP and with a stress on the non-deterministic mechanism. Another version will be developed on a Lisp Machine.

WEDNESDAY can analyze fairly complex, ambiguous sentences yielding the alternative interpretations. As an example consider the following Zen-like sentence, that has a number of different interpretations in Italian:

Il saggio orientale dice allo studente di parlare tacendo

WEDNESDAY gives all (and only) the correct interpretations, two of which are displayed in Fig.3a and Fig.3b (in English words, more or less: "the eastern treatise advices the student to talk without words" and "the oriental wisemen silently informs the student that he (the wiseman) is talking").

```
COGNITIVE NETWORK:
C0000183:
     P-BE-SILENT X0000175
C0000180:
     P-GER E0000178 C0000183
E0000178:
     P-TALK X0000175
C0000174:
     P-STUDENT X0000175
C0000165:
     P-ADVISE X0000076 E0000178 X0000175
C0000119:
     P-EASTERN X0000076
C0000075:
     P-TREATISE X0000076

     THREAD: C0000165
     (GAPS:)
- - - - - - - - - - - - - - W E D N E S D A Y
```

Fig. 3a

```
COGNITIVE NETWORK:
C0000245:
     P-BE-SILENT  X0000224
C0000242:
     P-GER  C0000225 C0000245
E0000240:
     P-TALK X0000224
C0000236:
     P-STUDENT  X0000237
C0000225:
     P-INFORM  X0000224  E0000240  X0000237
C0000223:
     P-ORIENTAL-MAN  X0000224
C0000217:
     P-WISEMAN  X0000224
     THREAD:  C0000225
     (GAPS:)
- - - - - - - - - - - - - - W E D N E S D A Y
```

Fig. 3b

Integration in WEDNESDAY of a mechanism for complex idiom recognition, taking into account different levels of flexibility that idioms display, is currently under development.

REFERENCES

Cappelli, A., Ferrari, G., Moretti, L., Prodanof, I. & Stock, O. 1978 An ATN parser for Italian: some experiments. Proceedings of the Seventh International Conference on Computational Linguistics (microfiche), Bergen.

Ferrari, G. & Stock, O. 1980 Strategy selection for an ATN syntactic parser. Proceedings of the 18th Meeting of the Association for Computational Linguistics, Philadelphia.

Hayes, P.J. & Mouradian, G.V. 1981 Flexible persing. American Journal of Computational Linguistic, 7, 232-242.

Kaplan, R. & Bresnan, J. 1982 Lexical-Functional Grammar: a formal system for grammatical representation. Bresnan, J., Ed. The Mental Representation of Grammatical Relations. The MIT Press, Cambridge, 173-281.

Kwansky, S.C. & Sondheimer, N.K. 1981 Relaxation techniques for parsing grammatical ill-formed input in natural language understanding systems. American Journal of Computational Linguistics, 7, 99-108.

Riesbeck, C.K. & Schank, R.C. 1976 Comprehension by computer: expectation-based analysis of sentence in context. (Research Report 78). New Haven: Department of Computer Science, Yale University.

Small, S. 1980 Word expert parsing: A theory of distributed word-based natural language understanding. (Technical Report TR-954 NSG-7253). Maryland: University of Maryland.

Stock, O. 1982 Parsing on WEDNESDAY: A Distributed Linguistic Knowledge Approach for Flexible Word Order Languages. (Technical Report 312). Roma: Istituto di Psicologia del Consiglio Nazionale delle Ricerche.

Weischedel, R.M. & Black, J. 1980 Responding intelligently to unparsable inputs. American Journal of Computational Linguistics, 6, 97-109.

Woods, W. 1970 Transition network grammars for natural language analysis. Communications of the Association for Computing Machinery, 13, 591-606.