Istituto per la Ricerca Scientifica e Tecnologica (IRST)
Povo - Trento, I-38050, ITALY

UNDERSTANDING MESSAGES IN A DIAGNOSTIC DOMAIN

FABIO CIRAVEGNA ¤
cirave@irst.it

¤ Most of the work reported in this paper was done when the author was working at CENTRO RICERCHE FIAT, Orbassano (Torino, Italy).

Abstract - The problem of coping with subject-matter sublanguages in text processing is well known in the natural language processing field. The main problem is to balance the use of generic knowledge sources and the specific needs of the sublanguages. This paper introduces the characteristics of the sublanguage found in diagnostic messages about automotive equipment failures, and discusses an architecture to analyse those messages. The model is based on a two-level partial parsing approach that uses a syntax-driven strategy to parse fragments of a sentence. A set of semantics-driven strategies is used to collapse the fragments. General knowledge sources are proposed for use as an independent syntax, a knowledge-based semantics, a pragmatic module and a two-level lexicon. The problem of balancing accuracy, robustness and efficiency in the message analysis is addressed. Finally some applied results are shown.

1. INTRODUCTION

Information extraction from text has started to gain market interest and a number of applications have arisen. Engelien & Mc Bride (1991) foresee a boom in the market before the end of the century. Particular interest has gained message processing (Sundheim (1991) and Sundheim (1992)). The demand of real text application caused a shift in NLP research focus: from weak and computationally intensive methods to task-driven methods (Jacobs & Rau (1993)). Specifically, an attempt was made to fill the gap between the efficient but imprecise classical information retrieval (IR) and the accurate but weak and computationally expensive NLP approaches. The first, which came mostly from the IR field, brought Conceptual IR models (CIR). These use a semantic description of the domain to guide the pattern matching operation, thereby overcoming most of the IR problems. The main drawback of CIR systems is that they are not able to cope with the real content of the text, as they just search for the information of interest. Despite this limitation, they were successfully applied both in commercial applications (Andersen & al (1992)), and in the MUC-4 conferences (Appelt & al (1993)).

The second approach came from the NLP field. Hobbs & al (1992) adopted a full text parsing approach integrated with a number of heuristics to get robustness and efficiency; unfortunately heuristic strategies failed, mainly because they were altogether inefficient (Appelt & al (1993)). Montgomery & al (1991) introduced partial parsing approaches to parse accurately only fragments of the input: sometimes a global analysis is tried, relaxing the requirement as soon as an ill-formed input is found. Other approaches integrate top-down and bottom-up strategies to accurately parse only some parts of the input (Jacobs & Rau (1989)); Mc Donald (1992) and Mellish & al (to be published) adopt some partial analysis of the input text using some limited agenda-based chart parsers. In any case, a number of focusing methods are used to insulate those parts and to help the analyser in parsing the input (Jacobs (1990)). The partial parsing approaches were successfully applied in real contexts (see for example Jacobs & Rau (1993) or Ciravegna, Campia & Colognese (1992)).

In the NLP field particular emphasis was given to the problem of transportability to new applications and domains through the use of pre-existing core knowledge sources, in particular syntactic grammars. Unfortunately the use of general knowledge sources stresses the **sublanguage** treatment problem in many applications. "A sublanguage is a language resulting from restriction on and deviation from the standard grammar of a natural language; often a sublanguage grows in a natural way through the use of the standard language, albeit in special circumstances" (Lehrberger (1986)). Among the different sublanguages, particularly interesting are what Harris defined as *subject-matter sublanguages*, because they are strictly related to some application or domain and are characterised by limited lexical, syntactic, semantic and discourse properties, and text structure properties. Two approaches were presented to cope with sublanguage deviations:

1. the hard-wiring of the sublanguage into the system; it gives the system efficiency and coverage, but limits its transportability; moreover, ad-hoc approaches, although efficient

on the deviant forms, introduce some risks on those parts of the input that do not belong to the sublanguage proper.

2. treatment of the deviation of the sublanguage as ill-formed input. Many techniques have been proposed recently: from semantics-driven approaches to NLP, to recovery strategies for syntax-driven methods. Semantics-driven techniques are robust in that they do not give any clear role to syntax: only some kind of relaxed syntactic information is spread throughout the semantic analyser or the semantic knowledge (see Hayes (1984) and Hahn (1989)) . Although they allow great robustness (Carbonell & Hayes (1984)), their main drawback is that most of the phenomena that are easy to cope with through the use of syntax, can sometimes be difficult to deal with (see Rullent & Poesio (1987) and Tomita & Carbonell (1987)). Moreover the syntactic information is spread throughout the lexicon or the analyser and is often difficult to maintain. In the syntax-driven approach some special techniques were introduced for reducing its weaknesses: syntax-driven recovery strategies (Mellish (1989)), semantics-driven recovery strategies (Kirtner & Lytinen (1991)), and the use of metarules (Weischedel & Sondheimer (1983)). All these approaches have advantages and disadvantages (see Kirtner & Lytinen (1991)), but - when used in a restricted domain - they are mainly inefficient, because they treat all the deviant forms of the sublanguage as ill-formed input, i.e. as exceptions; moreover they introduce many ambiguities in resolving the ill-formedness that could be easily reduced by the use of sublanguage principles. Even if their advantage is to allow great transportability, they do not take into account what Dunham (1986) pointed out for medical diagnostic statements, i.e. that "there are some sublanguages in which syntax plays the weak sister of the pragmatically based discourse rules in computing a semantic representation".

This paper presents an architecture for analysing diagnostic messages about automotive equipment failures. The basic idea is to use a partial parsing approach relying on general knowledge sources, reducing the use of hard-wired sublanguage features to the minimum. The appropriateness of the architecture will be evaluated on three levels: accuracy, robustness and efficiency. In the next section the main features of an Italian diagnostic automotive sublanguage will be presented.

2. THE DIAGNOSTIC SUBLANGUAGE

A set of diagnostic messages about automotive equipment failures written in Italian was analysed. Those messages were produced as routine work by the diagnostic personnel repairing the failures on prototype cars. These messages were composed by a simple description of a generic malfunctioning reported by a driver (for example: "Internal carpet gets wet when it rains"), followed by a first diagnosis written by a repairman (for ex. "the problem is caused by water seepage from the left door due to damage to the weather seal"), and followed by a final diagnosis analysing in great detail the problem (i.e.: "the damage was caused by ... "). A typical diagnostic text is shown in Figure 1. A message is identified by a reference number; a list of structured data shows the characteristics of the car (i.e. licence plate number, model, mileage, etc.); a descriptive text illustrates the main fault found by the driver; a first diagnosis (empty in the example) and a diagnosis (which is written by two different people in this case) complete the text.

Usually the messages are short (1-10 lines), written by different people (at least a driver and a repairman), agrammatical (the use of telegraphic language, loss of agreement between constituents, typing errors, use of jargon), cryptic and without cohesion (for the presence of implicit diagnostic knowledge). The domain is characterised by the presence of a great number of objects: about 8.000 car parts, 500 faults, 500 other objects, and so on. The domain knowledge on the possible events is shallow as it is not possible - given the current knowledge base engineering techniques and costs - to create a general diagnostic knowledge base on faults, because a car is too complex and short-lived a system to model; moreover any manufacturer produces a number of different models, and a different version of the same car model may mount different equipment (leading to different malfunctions and so on).

The intended goal was to extract the diagnostic information (main fault, chain of causes, chain of effects, car parts involved, etc.) from the text and build its semantic representation (illustrated in Figure 2).

The main linguistic features of the sublanguage are:

- many sentences are composed by just a complex noun phrase ("loss of screw for fastening gear to main bearing due-to[1] bad mounting of screw[2]");
- many NPs often lack prepositions and articles ("damage weather seal edge right front door" vs. "damage *to the* edge *of the* weather seal *on the* right front door"); the missing preposition is generally "di" (English: "of"), but "su" (on), "in" and "per" (for) are also frequent; the most important effect of that loss is the introduction of syntactic and semantic ambiguities (as in the Noun+Adjective+Noun construction that will be discussed later);
- the complex descriptions of the domain (mainly the car part descriptions) behave as nominal compounds: many syntactic behaviours (for example the rules for adjectives) are different from the standard rules of the language when these objects are present;
- the diagnostic descriptions are affected by the presence of linguistic modifiers (adjectival and adverbial phrases, negations, etc.) that often introduce subjective evaluations in quantity and/or quality that must be formalized in some way; an example is the form *not very correct mounting* ;

[1] In Italian the form "due to" is expressed through the preposition "per".

[2] As we mentioned before a sublanguage of the Italian was analysed. When in this paper an English text will be shown, a literal translation is to be intended. Note that the only way to build a nominal compound in Italian is through a structure analogous to the English "of-the"/"on the" construction. So in the examples a compound like "screw gear" is to be intended as "the screw of the gear" but where the loss of preposition and articles is signalled. We will try to reproduce the same incorrectness in the English text too, but some times it won't be easy.

such modifiers can contribute to highlighting a relation, describing an undesired object feature or even changing the semantic identity of the term they are applied to (Ciravegna & Giorda (1993));

- the sublanguage often contains some material that does not belong to the sublanguage proper; those parts may be whole sentences ("the part was sent to laboratory 123A for further tests" or "the car was parked for two days near a building yard") or just chunks ("an oil leak was found on *the car used for the press shows*"); these parts are generally not closely related to the diagnosis itself;
- text coherence is often affected by the presence of implicit knowledge ("car broken down; oil sump leaking"); in other cases the coherence is maintained through the use of nominal anaphors: references may be made both through the repetition of the head of the noun ("... leak of oil... . *The oil* ..."), via hypheronimy, or part-of relation ("seepage on the left door; *the weather seal* is damaged");
- a diagnosis may involve even small parts of a small component and knowledge about the domain objects is not precise about that ("the *trade-mark on the screw* was stamped in the wrong place causing a fracture of the screw itself");
- knowledge acquisition is a crucial point for an NLP system as Jacobs, Rau & Zernik (1991) pointed out; in the diagnostic domain, new descriptions or names may be introduced by new equipment or by the evolution of the sublanguage and must be acquired automatically.

The next section describes the basic architecture.

3. THE BASIC ARCHITECTURE

The architecture aims to balance three characteristics in analysing the diagnostic messages: efficiency, robustness and accuracy. **Efficiency** is needed because the diagnostic messages are texts composed of many sentences, often containing a complex terminology, introducing syntactic and semantic ambiguities; efficiency is particularly important when the system must operate in real time. **Robustness** is needed because the main features of sublanguages are the use of idiosyncratic forms (unusual syntactic rules), the presence of implicit knowledge (ellipsis or jargon) and the use of extra-grammatical language (Liddy & al (1991)). **Accuracy** is needed because the technical descriptions are precise in their content (being written by specialists), and the object descriptions they involve are complex from a linguistic point of view (they may be composed of up to ten words); moreover those descriptions are often very similar: roughly speaking accuracy must allow a distinction between *a crack on the weather seal of the ring of the left door* and *a crack on the ring of the weather seal of the left door* . On the other hand these descriptions are often affected by the deviations of the sublanguage, and accuracy allows resolution of those deviations by using not only knowledge of the world, but linguistic information, too.

The efficiency, robustness and accuracy requirements (although differently formulated and motivated) correspond more or less to the current issues in message processing; this characteristic was especially stressed by the MUC conferences (Sundheim (1991) and Sundheim (1992)).

The general architecture is presented in Figure 3. The input text is first processed by a **pre-processor**; after that comes a morphological analysis, which is used to recognise some kind of pattern, such as ranges of data or numbers, measurements, chemical formulas, etc. The pre-processor is based on a context free grammar; a dictionary of about 9.000 lemmata is used; the entries are not limited to domain words. The rest of the analysis consists of a sequence of steps performed once for each sentence in the text.

The first step (**object-recognition**) processes the sentence from left to right in a syntax-driven way to recognise some aggregates as NPs, PPs and the verbal groups. The analysis, although syntax-driven, is controlled by the semantic modules in order to limit recognition to groups of words that contribute in building semantically uniform aggregates (for example recognising the car parts, failures and test descriptions contained in a sentence). To do that three knowledge sources are used: a dependency grammar based syntax (based on about 150 production rules), a case frame representation of the domain (about 50 classes), and a phrasal lexicon in which the syntactic and semantic representations of all known objects are contained (about 8.000 car parts, 500 faults, 500 other descriptions).

The second step (**the skimmer**) decides whether each sentence contains diagnostically interesting content or not. It is a semantic function mapping from the semantic type of the case frames (associated to each object produced by the previous step), to a numerical value. If the numerical value produced by the sentence is greater than a heuristic threshold, the analysis of the sentence is continued; otherwise it is interrupted and the sentence content is left only for solving some anaphoric references.

When a sentence is accepted by the skimmer, a connection among the objects contained is tried, using some semantic-driven strategies (**object-linking**). There are different kinds of connections depending on the level of recognition reached in the previous step, as will be seen later.

The **interpretation** is used to join the information brought by the current sentence to the global meaning of the text. The cohesion of the text is maintained mainly through the resolution of the anaphoric references. During this pass a second run of the skimmer eliminates the uninteresting sentences that were not filtered by the previous run.

Finally, a **template** is filled through a normalisation of the diagnostic knowledge extracted from the text. A set of heuristic rules groups is adopted. They normalise the chain of causes/effects into a chain of causes only. For example, in a text such as that in Figure 1, there is a chain of causes composed of a fault, its cause, the two effects of the causes and a cause of the cause. During this pass the chain illustrated in Figure 2 is generated. The information extracted from the text that is not relevant for the template filling is ignored. In the rest of this section some of the details of the architecture illustrated above are discussed.

3.1 OBJECT RECOGNITION

The object recognition step processes the sentence from left to right in a syntax-driven way to recognise some aggregates as NPs, PPs and verbal groups. For example, given the simple description:

"a loss of the screw for fastening the gear to the main bearing was found" the object-recognition role is to recognise three object descriptions: a fault ("the loss"), a car part ("the screw for fastening the gear to the main bearing") and a verbal group ("was found"). Recognising an object means assigning a syntactic structure, a case frame representation, and a search code in a data base (if present).

Three knowledge sources are used at this level: lexical, syntactic and lexical-semantic knowledge. The role of syntax is to give a linguistic structure to the NPs, PPs and verbal groups that are used to introduce these objects. The role of the lexicon is to give information about the words and their aggregates (including the names of the known car parts, failures, etc.[3]); the semantic module gives the description of the domain (through a case frame taxonomy).

Syntactic knowledge is represented by a formal grammar on which an independent module operates. The initial model was derived from that used in FIDO (Lesmo & Torasso (1985)) and in a first version of SINTESI (Campia & Colognese (1990)). The syntactic analysis is based on a dependency grammar containing about 150 rules. The main feature of such a tree with respect to the normal trees used for constituency based grammars is that every node contains a word. A dependency tree for the sentence "the man kicked the red ball" is presented in Figure 4. Syntactic analysis is done by a set of production rules (IF <conditions> THEN <actions> rules). The conditions test the current tree status, whereas the actions modify it. There are three groups of rules (Lesmo & Torasso (1985)):

1. Construction/Enlargement rules: they enlarge the dependency tree starting from the current tree status.
2. Semantic and agreement control rules: every enlargement must be tested by the semantic module; it is done immediately to avoid the explosion of possibility due to structural ambiguity, and to have a deterministic parser.
3. Natural change rules: a method to modify the tree if the semantic tests fail.

The analysis is deterministic; the only backtracking allowed is on actions operated on the current node (=word) and those introduced by the natural change rules.

The semantic analyser runs in parallel with the syntactic module: each semantic control rule (group 2 above) activates the semantic module to test the connection. The semantic module builds the semantic representation of each object using two types of knowledge sources: a domain model represented by case frames (about 50 classes), and a phrasal lexicon. The case frame model was derived from the Entity Oriented Parsing proposed by Hayes (1984). A case frame is activated for each object; it models the behaviour of the object in the domain.

The phrasal lexicon is composed of structured syntactic trees grouped by using semantic criteria (about 8.000 car parts, 500 faults, 500 other descriptions). The entry point is the syntactic head of each

[3] The descriptions of the objects of the domain are stored in a phrasal lexicon; in some previous papers they were referred to as *deep semantic knowledge* .

description. Object descriptions and case frames are activated by the syntactic module when the object head is found, and they constitute a first hint of what could follow the head in the sentence.

For example, let us analyse the simple description: "Fissaggi della coppa dell'olio del motore con cricche" (literally “Bolts of the sump of the motor oil with cracks”). First of all syntax, finding the syntactic (and semantic) head of the first object (“fissaggi”), requires the activation of a new empty syntactic tree and asks the semantic module to activate the semantic descriptions of the object (i.e. a case frame *car-part* and all the *bolt* descriptions). Some of these descriptions are illustrated in Figure 5; the corresponding trees are in Figure 6.

The syntactic module will connect “della” (“of the”) + “coppa” (“sump”) and then “fissaggi” (“bolts”) and “coppa” (“sump”); each connection is immediately checked by the semantic module; the semantic test is performed using the information associated to each word in the lexicon, and comparing the current syntactic tree with the available descriptions in Figure 6. An efficient algorithm was developed for the comparison, because of the great number and complexity of similar objects in the domain (for example there exist about 300 different gaskets whose descriptions are very similar); the efficiency is necessary in particular to cope with the ellipses caused by the conjunctions (Ciravegna, Campia & Colognese (1991)).

Articles and prepositions, although not present in Figure 6, are checked by the lexical-semantic module in a different way: they are considered optional and, when present, must represent relations acceptable for the kind of description they are applied to; for example in a car part description they must represent *locations* or *part-of* relations. In case of deviation for the default, it is possible to put the preposition in the description, as was done for the “for” preposition in the last tree in Figure 6. The current object is considered completed when the next word is not part of the current object (for semantic, syntactic or pragmatic reasons); in the example above, the end of the car-part is recognised when the word “cricche” (“cracks”) is found. When an object is completed, the syntactic tree and the semantic descriptions are stored in a queue. The result of the analysis at this step is an ordered collection of syntactic trees, case frames and semantic descriptions representing the objects of the sentence.

3.1.1 Accuracy, robustness and efficiency in object recognition

During the object-recognition step, the role of syntax is to give a linguistic structure to the NPs, PPs and verbal groups that describe the objects of the domain (car parts, failure, etc.). The use of a general grammar guarantees the ability of coping with most of the general linguistic phenomena that are present at this level. **Accuracy** is guaranteed by the syntax-driven analysis controlled both by the semantic module and the phrasal lexicon. The approach is **efficient** too because the strict syntax-semantics interaction allows easy reduction of the overhead caused by the PP-attachments that are very common in syntax-driven approaches. Moreover, the analysis is deterministic.

The deviation in regard to standard language is mainly represented at this level by the **lack of prepositions and articles** in the descriptions. It is possible to overcome this problem because the use of head and modifier structures and rules allows the production rules to be transformed into operations on graphs limited to the current tree situation and the syntactic type of the current word. It is easier to cope with ill-formed input in this way (Campia & Colognese (1990)). Robustness, efficiency and accuracy were particularly useful in resolving ellipses caused by the conjunctions (Ciravegna, Campia & Colognese (1991)).

An interesting case of ill-formed input is the **"Noun + Adjective + Noun" sequence**: the latter is typical of Italian telegraphic language, in which the adjective may refer to both the nouns because the preposition for the second noun is lost. It is necessary to perform semantic controls and tests of agreement (in number and gender) before deciding the right connection. It is a syntactic task to know how to tackle this task, asking semantics to execute some look-ahead action, for understanding the meaning and role of both the nouns before testing the adj+noun2 connection; in fact in a description such as: “perdita grave olio cambio" [literally :"spill heavy oil gear"] in which "olio" is referring to another semantic object with respect to "perdita" (contained in the current tree), it is necessary to instantiate the “olio” object (with its case frame and associated descriptions) before testing the "grave” + “olio" connection. A better description of this point may be found in (Campia & Colognese (1990)). This is a case in which the sublanguage features had to be hard-wired into the system; anyway the Noun + Adjective + Noun treatment constitutes a packet of rules separate from the other parts of the grammar and easily eliminated when necessary (i.e. for an application without such possibilities).

The idea of limiting the parsing to some sections to avoid the combinatorics of parsing was present in many previous works: in Mc Donald (1992) the text is pre-segmented to aid a chart-based parser; Hobbs & al 1992 proposed a segmentation to reduce the combinatorics of parsing in long sentences only. The most interesting approach is probably in Jacobs (1990) and Jacobs, Krupka & Rau (1991) in which a skimmer is proposed to tag each word, segment the texts and pre-process the input to assign the duty of connecting some constituents to tasks in which world knowledge plays a large role. This paper presents an approach that takes into account Jacobs’ proposals to use not only morphological hints to segment the text, but semantics too; the main difference is that while the other authors completely avoided parsing before skimming, here the parser itself is used to segment the text, and the skimmer is applied afterwards. It is necessary because even a sophisticated skimmer operating on the plain text could be fooled by some diagnostic descriptions: for example a car part description such as “spia per mancanza di olio” (led for/due-to[4] the lack of oil”) could be confused by the skimmer with a description of a car part (“led”) and a fault (“lack of oil”); operating after the recognition of

[4] In Italian the preposition “per” is ambiguous between “for” and “due-to”.

object is useful for recognising these cases, which are very frequent. On the other hand the object recognition step is efficient enough that it does not slow down the analysis too much; moreover the rate of the uninteresting sentence in the diagnostic domain doesn't exceed 10% of all sentences (against the 90% of the MUC-4 domain), and a small loss in efficiency gives better precision. Other cases in which high accuracy is required are descriptions such as "funzionamento non molto corretto" ("not very correct working"): they must be analysed very carefully to correctly understand their real diagnostic content (Ciravegna & Giorda (1993)). Moreover an at least partially parsed sentence is necessary to cope with anaphoric references in the subsequent sentences, especially for part-of relations.

3.2 OBJECT LINKING

When a sentence is accepted by the skimmer, a connection among the objects contained is tried (**object-linking**). There are two possible kinds of connections, depending on the level of recognition achieved in the previous steps: total linking and partial linking. The first is used to build a complete semantic description of the sentence, and is tried when the object recogniser reports a high reliability rate in its results (depending for example on the number of unknown words, etc.). It integrates Bottom-Up (BUS) and Top-Down (TDS) semantic strategies: the **TDS** is an expectation-driven analysis of the connections among objects, driven by the main roles of some constituents (for example the agent and the direct object of a verb, the roles of a conjunction, etc.). Each TDS connection is tested by syntactic and pragmatic rules. The TDS is not left-to-right. The **BUS** considers the object from left to right instead, connecting the objects that are still dangling after the TDS step. For each object a role-driven connection is tried with all those objects that are linguistically and semantically acceptable. Syntactic and pragmatics rules check each connection. The way the system operates on the ill-formed sentence "Durante montaggio perno rivela rottura di testa e fungo" (lit. "during mounting hub presents fracture of head and mushroom") is illustrated in Figure 7: the edges represent the actions, the numbers the order of connection; the TDS connections are shown above the sentence, the BUS below.

For a sentence analysis with low reliability, only the **BUS** is adopted to form some aggregates of objects. This aggregation is also influenced by the presence of obscure parts. Connections among concepts separated by such parts are considered unlikely. A classification of the obscure parts is made trying to apply some lexical, syntactic or semantic heuristic rules to the identity of the words contained (a verb has a strong power of separation, a group of adjectives has not, and so on).

Note that even the TDS+BUS allows the system to cope with sentences that do not arrive at a complete structure because the BUS guarantees the robustness.

3.2.1 Accuracy, robustness and efficiency in object linking

Object linking is mainly driven by the semantic module; the role of syntax at this level is only to test the semantic choices, i.e. it "plays the weak sister of the pragmatically based discourse rules in computing a semantic representation" as Dunham (1989) pointed out. The important features obtained with this approach are robustness and efficiency. The approach is **robust** because it has all the potentiality of the semantic approaches to cope with ill-formed input (Carbonell & Hayes (1984)); moreover it uses its semantic previsional ability to cope with unknown parts of the input. A special kind of unknown input is the presence of unknown technical object descriptions that are very frequent in a diagnostic message; as a matter of fact a diagnosis may involve even small parts of a small component, and the domain description is not precise about that ("errato stampaggio *marca vite...* causa rottura vite stessa" i.e. "Bad printing of *trade-mark screw...* caused fracture screw itself"); recognising those parts is not easy, especially because the lexicon is not reliable for that; the object recognition segments them into a sequence of objects using the lexicon (in the example "trade-mark" + "screw..."). The semantic module in the object linking step is able to understand (using the semantic expectations) that it is just one description of an unknown object. The two objects are hence collapsed in a unique description and the system manager then can be asked for (optional) insertion into the phrasal lexicon. The ability of recognising new descriptions is crucial for the **transportability** to new applications: it allows the object descriptions to be built in a semi-automatic way. Moreover it is an interesting feature in the automotive diagnostic domain because new descriptions or names may be introduced by new equipment or by the evolution of the sublanguage and these can be acquired semi-automatically. The reliability rate of this strategy exceeded 99% in a test on about 1.000 messages, in which about two unknown descriptions per text were found.

This approach is able to cope with some kinds of metonymy. A typical case is the following: "A loss of the screw for fastening the hub was found with a subsequent loss of the hub itself. *The hub* damages the gasket". In diagnostic terms it is not the hub itself that causes the damage to the gasket, but *the loss* of the hub. The semantic module (both during TD or BU strategies) is able to solve the metonymy using both the semantic expectation given by general world knowledge and the information given by the pragmatic interpreter that resolves the anaphoric reference.

The approach is **efficient** because it relies on heuristic semantic strategies to reduce the overhead involved by PP-attachments (BUS). This is a crucial characteristic in a domain in which a sentence may contain up to 10 PPs in sequence, as in:

> "mancato funzionamento del motorino avviamento durante prova pergola per ossidazione con conseguente bloccaggio innesto alberino scorrimento, e conseguente mancata chiusura contatti elettromagnete per l'utilizzo di materiale non idoneo alle prescrizioni"

Another point for **efficiency** of the approach is the ability to cope with some garden paths, frequent in the diagnostic sublanguage. For example in Figure 7 an ambiguity arose in the resolution of the gap between "mounting" and "hub" (in Italian *of the* is just one word, so both the corrections [the/of-the] are plausible in the same way). A preference for the "of-the" correction (plausible for the near presence of the "mounting" object) could lead to a garden path. The approach avoids the garden path because it applies first the TD strategy, which looks for the main roles of the verb (leading to the "the" correction in the previous example), and then the BUS for the other roles (such as PP-attachments).

TDS and BUS are cases in which the characteristics of the sublanguage are hard-wired into the system. Moreover the method is semantic-driven and could seem an unwise choice, because many of the linguistic phenomena that are easy to cope with for a syntax-driven method (for example the comparatives) are difficult to treat; many of the drawbacks of the true semantic approaches were reported. Anyway it must be noted that:

- the sentences of the text analysed were in general very simple from a syntactic point of view, but complex from a semantic and terminological one;
- each object connection required some degree of robustness: from the loss of prepositions and articles, to the presence of some unknown object descriptions (split in many different descriptions by the object recognition step), to the presence of unknown words;
- a great overhead in the PP attachment was reported.

It was then necessary to privilege robustness and efficiency more than linguistic precision at this step. The linguistic simplicity of the sentence allowed a high degree of precision to be obtained with a semantics-driven approach just controlled by the syntactic module; unfortunately some simple syntactic features about how a sentence may be formed had to be in some way hard-wired into the semantic processor and are then duplicated in the syntactic grammar. The semantic processor asks syntax for the greatest part of syntactic information, so the duplication is not too heavy.

As was mentioned before, the idea of using world knowledge to help the connections among objects of the sentence is not new: in Jacobs & Rau (1989) and Jacobs & Rau (1993) top-down strategies are proposed that use some score preferences to decide which is the best connection, in (Weischedel (1991)) a stochastic parser generates fragments that are collapsed by a semantic interpreter and a discourse processor. In the proposed approach both BUS and TDS use score preferences based on the knowledge of the world, lexical semantics and syntax.

3.3 INTERPRETATION

The interpretation is used to join the information brought by the current sentence to the global meaning of the text. Text cohesion is maintained mainly through the resolution of the anaphoric references. There are two types of references that are treated: nominal anaphora and part-of references. Two kinds of nominal anaphors are recognised: via hyponimy/hypheronimy relations (i.e. "...a fracture... . The **failure**...") and via (partial) repetitions of the phrases (i.e. "...a fracture... . The **fracture** ..."). The referenced object is retrieved during the object recognition step through a search guided by the syntactic structure of the previous sentences of the text. The model was derived from that presented by Allen (1987). The structure of the text as a whole is not considered, as the messages coped with were very simple from this point of view.

The *part-of reference* is particularly interesting in such a domain because it mainly affects the car part descriptions and is very frequent. A typical text contains a precise reference to the object that presents the main fault in its introduction (the description of the main fault). All the other objects in the texts are related to that precise description. For example a typical text is presented in Figure 1: in the first sentence the object "starter motor" is mentioned, and in the rest of the message some of its parts are mentioned ("starter drive pinion" and "electromagnetic contacts"; the complete names are "starter drive pinion of the starter motor" and "electromagnetic contacts of the starter motor"). It is a part-of reference in which the part-of relation is left implicit by an ellipsis in the descriptions[5]. Those ellipses are to be solved without relying on the domain described by the case frame taxonomy, because - as was pointed out previously - it is not precise about the model of the car system. The only way to operate is by relying on the lexicon, or, better, on that part of the lexicon that contains the object descriptions. Actually the only way to resolve the gaps is to compare all the possible descriptions activated by "starter drive pinion" and "electromagnetic contacts" during the object recognition step, with the "starter motor" description. It means comparing the syntactic trees representing the actual object description (Figure 8, top right) with all possible descriptions (Figure 8 on the left, shows two) using the other descriptions of the objects in the whole text (Figure 8, right-bottom).

Accuracy and efficiency are guaranteed by a characteristic of the comparison algorithm - it takes into account the syntactic structure of the descriptions (Ciravegna, Campia & Colognese (1991)) and by the retrieval algorithm that takes into account the structure of the text. Note that efficiency and accuracy are crucial during this pass because:

- the reference can be underspecified (just one word as in "the gasket");
- there can be many different objects in the sentence and only one of them is to be chosen for resolution;
- the number of possible descriptions to compare against can be high (for example each object can have associated to it up to about 300 different descriptions).

Note that the retrieval of the completed object gives preference to the current sentence (trying to look both up-ward and down-ward) and then to the rest of the text. A complete description of those algorithms can be found in Ciravegna (1989-1993).

4. CONCLUSION

This paper presented an experience in analysing diagnostic messages about automotive equipment failures written in Italian. Those messages are mainly short (1-10 lines), written by different people,

[5] This kind of references and ellipsis don't affect or involve the text cohesion, but the correct semantic understanding of the diagnosis.

agrammatical (because of the use of a sublanguage), cryptic and without cohesion. The domain is characterised by the presence of a great number of objects and shallow general world knowledge.

A two-level approach to parsing diagnostic messages was proposed. It adopts a syntax-driven strategy to parse fragments of sentences and a set of semantics-driven strategies to collapse the fragments. The basic idea is to use general knowledge sources and modules (i.e. an independent syntax, a knowledge based semantics, a pragmatic module and a two level lexicon), reducing the hard-wired sublanguage features to the minimum.

The two level architecture is identifiable as a partial parsing approach and is consistent with the current issues in message processing; in particular with the idea expressed by Jacobs & Rau (1989) and (Weischedel (1991)). The idea of limiting analysis to some parts of the input to avoid the combinatorics of parsing through the use of a kind of pattern matching (although differently formulated and motivated) is consistent with the suggestions made by Mc Donald (1992), Hobbs & al (1992), Jacobs (1990) and Jacobs & Rau (1993).

It was then demonstrated that it is possible to balance accuracy, robustness and efficiency. Accuracy is achieved because the syntax-semantics interaction during object recognition step and TDS+BUS during object linking lead to a full structural and semantic definition when the input is correct. Robustness is achieved during the object recognition step because the syntactic analyser is able to cope with some ill-formed input (the loss of prepositions and articles); it is achieved during the object linking step, adopting different strategies, leading to partial analysis when complete sentence analysis is impossible. Efficiency is guaranteed by the interleaved interaction syntax-semantics and by the semantics-driven strategies. The introduction of the skimmer brings both efficiency and robustness.

The use of an independent phrasal lexicon allows easy transportability through different applications in the same diagnostic domain (for example from car diagnosis to truck diagnosis). Moreover, transportability is favoured by the ability to build the phrasal lexicon in a semi-automatic way. This is an interesting feature in the diagnostic domain because new descriptions or names may be introduced by new equipment or by the evolution of the sublanguage, and they can be acquired automatically.

The hard-wired sublanguage features are limited to the treatment of some special syntactic forms (i.e. the Noun + Adjective + Noun form) and to the organization of the semantic strategies (BUS and TDS).

The proposed model was implemented in an applied system (Ciravegna, Campia & Colognese (1992)), (Ciravegna & Giorda (1993)). The system had a vocabulary of about 9.000 lemmata, a phrasal vocabulary of about 6.000 descriptions and a taxonomy of about 50 case frames describing the diagnostic domain. The grammar was composed of about 150 dependency rules. It was mainly written in C-language, using the Nexpert Object tool. It was tested on about 1.000 texts with 85% recall and 90% precision. It was able to process about 150 messages per hour. It was tested by some users too with analogous results. The role of the system was to populate a database and to build a knowledge base on faults.

5. ACKNOWLEDGEMENTS

The work reported here is the result of my personal experience of some years of research in knowledge extraction from texts, mostly spent as researcher at Centro Ricerche FIAT in Orbassano (Torino, Italy). This paper is the result of putting together the papers I already published before, in particular those that are listed in the bibliographical section. The results reported would not have been possible without the ideas and support of a number of people. I would like to thank Paolo Campia and Alberto Colognese for the development of the model for syntax-semantics interaction and for many discussions during two years of intense collaboration. The idea and development of the phrasal vocabulary came from discussion with L. Lesmo, P. Terenziani and V. Lombardo (from the University of Torino, Italy), and P. Campia, A. Colognese and R. Tarditi . The contribution of Erica Giorda to the model for the lexical modifiers was fundamental and is described in Ciravegna & Giorda (1993). I'm very grateful to Massimo Zancanaro for the discussion we had about text processing and for the help on terminology. Finally I would like to thank Oliviero Stock and Alberto Lavelli for their suggestions about this paper.

All the content of this paper must be considered my personal opinion and must not be interpreted as representing the ideas, the policies or activities on natural language of the FIAT Research Center. The diagnostic texts used in the examples are invented.

6. BIBLIOGRAPHY

Allen J. (1987). *Natural Language Understanding,* (pp. 334-365). Menlo Park, CA: The Benjamin Cummings Pub. Company.

Andersen P., Hayes P. J., Huettner A. K., Schmandt, L. M., Niremburg I. B. & Weinstein S. P. (1992, March). *Automatic Extraction of Facts from Press Releases to Generate News Stories.* Paper presented at the 3rd Conference on Applied Natural Language Processing, Trento, Italy.

Appelt D.E., Hobbs J.R., Bear J., Israel D. & Tyson M. (1993, August). *FASTUS: A Finite-state Processor for Information Extraction from Real-World Text.* Paper presented at the 13th International Joint Conference on Artificial Intelligence (IJCAI93). Chambery, France.

Campia, P. & Colognese, A. (1990, October). *Organizzazione della Conoscenza Sintattica e Interazione con la Semantica in un Sistema per la Comprensione di Testi.* Tesi di Laurea, Torino, Italy.

Carbonell, J.G. & Hayes, P.J. (1984). Recovery Strategies for Parsing Extragrammatical Language. *American Journal of Computational Linguistics* , 9 (3-4), 123-145.

Ciravegna F. & Giorda E. (1993, October). Coping with Modifiers in a Restricted Domain, in P. Torasso (Ed.) *Advances in Artificial Intelligence* (266-271), Lecture Notes in Artificial Intelligence (LNCS), Berlin, Germany: Springer-Verlag.

Ciravegna, F., Campia, P. & Colognese, A. (1991, October): The treatment of Conjunctions in an Information Retrieval System, in E. Sorbello (Ed.) *Le prospettive Industriali dell'Intelligenza Artificiale.* Palermo, Italy.

Ciravegna F. (1989-1993). SINTESI, 5 Technical Reports, Orbassano (Torino), Italy, 1989, 1990, 1991, 1992, 1993.

Ciravegna F., Campia P. & Colognese A. (1992, August). *Knowledge Extraction from Text by SINTESI*, Paper presented at the 14th Conference on Computational Linguistics (COLING92), Nantes, France.

Ciravegna, F., Tarditi, R., Campia, P. & Colognese, A. (1991, April). *Syntax and Semantics in a Text Interpretation System*; Paper presented at the RIAO91 Conference on Intelligent Text and Image Handling, Barcelona, Spain.

Ciravegna F. (1994, January). Estrazione di Informazioni da Testi: Stato dell'Arte e Prospettive, IRST Internal Report # 9401-05, Trento, Italy.

Dunham G. (1986). The Role of Syntax in the Sublanguage of Medical Diagnostic Statements, in R. Grishman and Kittredge (Eds.) *Analysing Language in Restricted Domains: Sublanguage descriptions and Processing*. Hillsdale NJ: Lawrence Erlbaum Associates, Publishers.

Engelien B. & Mc Bryde R. (1991, July): Natural Language Markets: Commercial Strategies, OVUM Edition.

Hayes, P.J. (1984, July). *Entity Oriented Parsing,* Paper presented at the 10th Conference on Computational Linguistics (COLING 84), Stanford, CA.

Hahn U. (1989): Making Understanders out of Parsers: Semantically Driven Parsing as a Key Concept for Realistic Text Understanding Applications, in Yager, R. (Ed.) *International Journal of Intelligent Systems,* 4 (3), 345-393.

Hobbs J. R, Appelt D. E., Bear J. & Tyson M. (1992, March). *Robust processing of Real-World Natural Language Texts.* Paper presented at the 3rd Conference on Applied Natural Language Processing, Trento, Italy.

Kirtner J. D. & Lytinen S. L. (1991). *ULINK: A Semantics-Driven Approach to Understanding Ungrammatical Input.* Paper presented at the 9th Conference of the American Association for Artificial Intelligence (AAAI), Anaheim, 1991.

Jacobs P. S. & Rau L. F. (1989, February). *Integrating Top-Down and Bottom-Up Strategies in a Text Processing System.* Paper presented at the 2nd Conference on Applied Natural Language Processing, Austin-Marriot at the Capitol, TX.

Jacobs P.S. (1990, August). *To Parse or Not to Parse: Relation-Driven Text Skimming.* Paper presented at the 13th Conference on Computational Linguistics (COLING90), Helsinki, Finland.

Jacobs P. S. & Rau L. (1993). Innovation in Text Processing, *Artificial Intelligence* , 63 (1-2), 143-191.

Rau L. F., Jacobs P. S. & Zernik U. (1989). Information Extraction and Text Summarisation Using Linguistic Knowledge Acquisition. *Information Processing and Management,* 25 (4), 419-428.

Jacobs P.S., Krupka G.R. & Rau L.F. (1991, February). *Lexico-Semantic Pattern Matching as a Companion for Parsing in Text Understanding,* Paper presented at the Speech and Natural Language Workshop, Pacific Grove, CA.

Lehrberger J. (1986). Sublanguage Analysis. In R. Grishman and Kittredge (Eds.) *Analysing Language in Restricted Domains: Sublanguage descriptions and Processing,* Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Lesmo L., Torasso P. (1985). *Weighted Interaction between Syntax and Semantics in Natural Language Analysis.* Paper presented at the 9th International Joint Conference on Artificial Intelligence (IJCAI 85), Los Angeles, CA.

Liddy E.D., Joergenson C.L., Sibert E. & Yu E.S. (1991, April). *Sublanguage Grammar in Natural Language Processing for an Expert System.* Paper presented at the RIAO91 Conference on Intelligent Text and Image Handling, Barcelona, Spain.

Mc Donald D. D. (1992, March). *An Efficient Chart-based Algorithm for Partial-Parsing of Unrestricted Texts.* Paper presented at the 3rd Conf. on Applied Natural Language Processing, Trento, Italy.

Mellish C. (1989, June). *Some chart based techniques for Parsing Ill-formed Input.* Paper presented at the 27th Annual Meeting of the Association for of Computational Linguistics, Vancouver, BC.

Mellish C., Allport D., Hartley A.F., Evans R., Cahill L.J., Gaizauskas R. & Walker J.: The TIC Message Analyser, manuscripted paper.

Montgomery C.A., Glover Stalls B., Belvin R.S. & Stumberger R.E. (1991, May). Language Systems Inc.: Description of the DBG System as Used for MUC-3 . In Sundheim, B. (ed.) *Prooceedings of the 3rd Message Understanding Conference (MUC-3).* San Diego, CA. (Distribution: San Mateo California: Morgan Kaufmann Publishers Inc.)

Rullent, C. & Poesio M. (1987, August). *Modified Case frame Parsing for Speech Understanding* .Paper presented at the 10th International Joint Conference on Artificial Intelligence (IJCAI 87). Milano, Italy.

Sundheim, B. (Ed.) (1991, May). *Prooceedings of the 3rd Message Understanding Conference (MUC-3).* San Diego, CA. (distribution: San Mateo California: Morgan Kaufmann Publishers Inc.)

Sundheim, B. (ed.) (1992, June): *Prooceedings of the 4th Message Understanding Conference (MUC-4).* McLean, Virginia. (Distr. San Mateo California: Morgan Kaufmann Publishers Inc.).

Tomita, M. & Carbonell J. (1987, August). *Another Stride towards Knowledge-Based Translation.* Paper presented at the 10th International Joint Conference on Artificial Intelligence (IJCAI 87). Milano, Italy.

Weischedel R. M. & Sondheimer N. D. (1983). Metarules as a Basis for Processing Ill-Formed Input. *American Journal of Computational Linguistics* , 9 (3-4), 161-177.

Weischedel R., Ayuso D., Boisen S., Ingria R. & Palmucci J.: BBN: Description of the PLUM system as used for MUC-3. In Sundheim, B. (ed.) *Prooceedings of the 3rd Message Understanding Conference (MUC-3).* San Diego, CA. (Distribution: San Mateo California: Morgan Kaufmann Publishers Inc.)

REF.: 00140/89
STRUCTURED DATA: <licence plate number, model, km, ...>
TOPIC: Mancato funzionamento motorino avviamento.
TEXT: Sulle auto per presentazione a stampa specializzata si verifica il mancato funzionamento del motorino avviamento durante prova pergola (motorino EY8 0, 8/72).
FIRST DIAGNOSIS: Antonioli 24/06/89: vedere scheda 0014/89.
DIAGNOSIS: Bianchi 25/06/89: Anomalia causata da ossidazione con conseguente bloccaggio innesto alberino scorrimento, e mancata chiusura contatti elettromagnete. Il particolare □ stato inviato ai laboratori per ulteriori controlli.
Giorgioni 28/06/89 l'ossidazione e' stata causata dall'utilizzo di materiale non idoneo alle prescrizioni.

TEMPTATIVE ENGLISH TRANSLATION:
REF.: 00140/89
STRUCTURED DATA: <licence plate number, model, mileage, ...>
TOPIC: Non-functioning motor starter .
TEXT: on the cars for presentation to the specialised press the motor starter didn't work during a test at the booth (motor EY8 0, 8/72).
FIRST DIAGNOSIS: Antonioli 24/06/89: see data sheet 0014/89.
DIAGNOSIS: Bianchi 25/06/89: anomaly caused by oxidation with consequent blockage starter drive pinion , and failure to close electromagnetic contacts. The part was sent to the laboratories for further tests.
Giorgioni 28/06/89 oxidation was caused by use of material not suitable to necessity.

Figure 1. Problem identification and resolution text.

```
Main Fault: non-functional  (cod. A124)
        Part: starter motor  (cod: 0129AIX2)
        During: booth test   (cod. P128)
        Caused by:  failure to close (cod. A156)
            Part: electromagnetic contacts starter motor (cod 0129OOT9)
            Caused by: blockage  (cod A345)
                  Part: starter drive pinion (cod. 0129OOT9)
                  Caused by: oxidation (cod A567)
                     Part: starter drive pinion (cod. 0129OOT9)
                         Caused by: use of unsuitable material  (cod. A178)
                                  Part: starter drive pinion      (cod. 0129OOT9)
```

Figure 2. Normalised chain of causes.

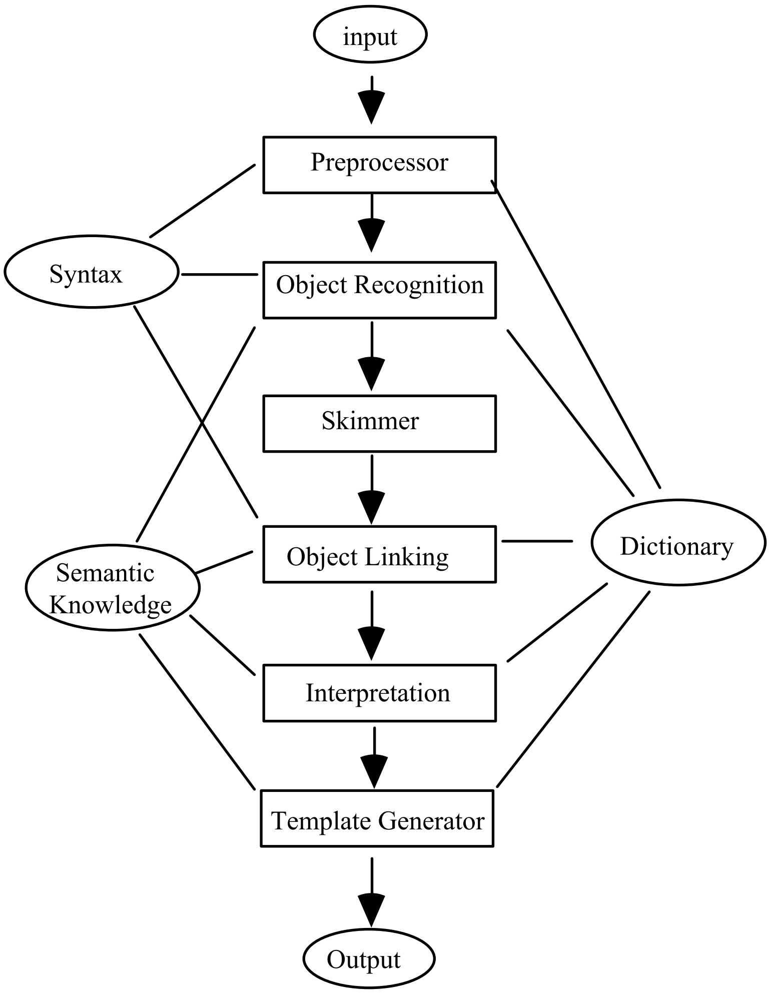

Figure 3: The general Architecturε

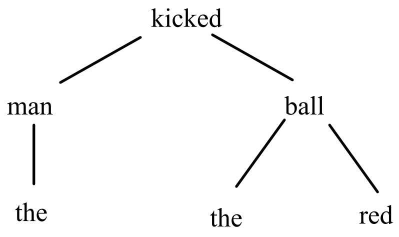

Figure 4: Sample Dependency Tree.

a. (bolt (hub (block (motor))))	code A123
b. (bolt (sump (oil (motor))))	code A342
c. (bolt (shock absorbers (front (frame))))	code X87C
d. (bolt (sump (to recycle (oil))))	code W09R

Figure 5: Sample Descriptions

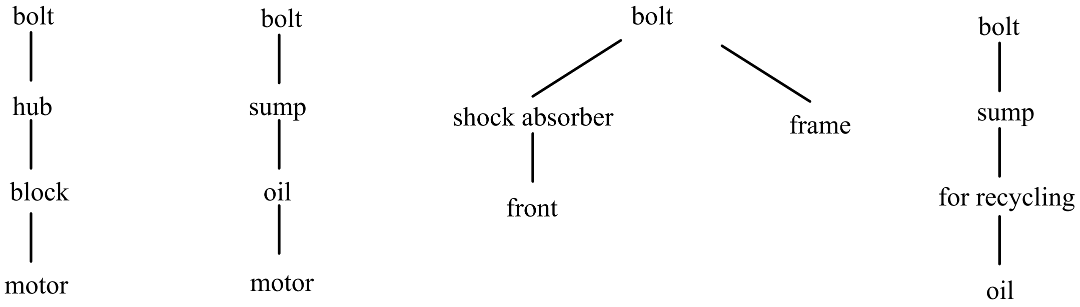

figure 6: Sample Syntactic Tree

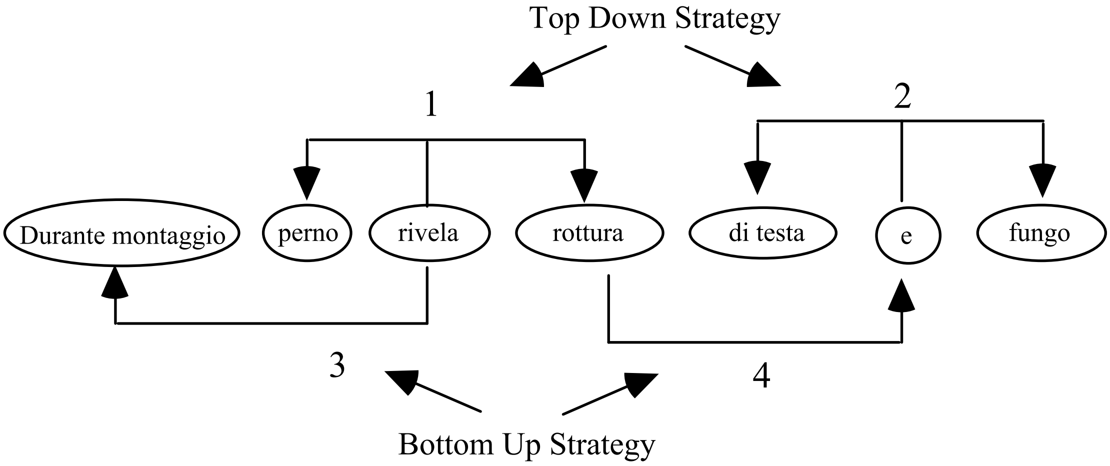

Figure 7: Strategies for: "During mounting hub presents break of head and mushroom'

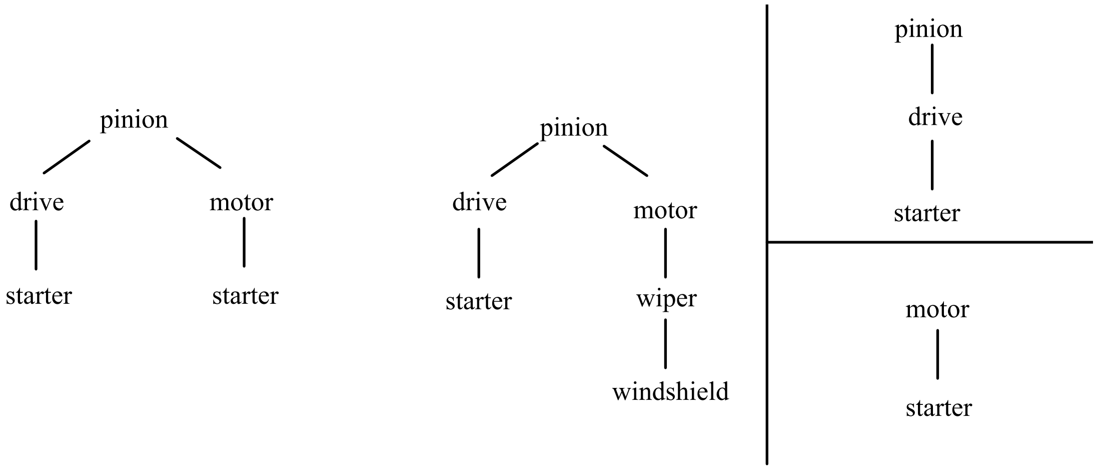
pinion
drive
motor
starter
starter
pinion
drive
motor
starter
wiper
windshield
pinion
drive
starter
motor
starter

Figure 8.