Simonetta Montemagni,
Dipartimento di Linguistica, Università di Pisa

Tailoring a broad coverage grammar for the analysis of dictionary definitions

ABSTRACT: Syntactic analysis is the first step of a two-stage strategy for extracting semantic information from dictionary definitions. By parsing definitions with a broad coverage Italian grammar, the distinguishing properties of dictionary language can be discovered, namely, greatly reduced ambiguity and otherwise ill-formed input which occurs regularly in definition texts. These properties have been exploited for tailoring an existing parsing system to meet the needs of dictionary text type; a post-processor which modifies and improves the parses produced by the general grammar has been added. This refinement is illustrated by means of examples.

1. Introduction

Dictionaries are rich sources of detailed semantic knowledge expressed in Natural Language (NL). Machine readable dictionaries are being exploited by automatically extracting that semantic knowledge from the dictionary definitions in order to construct lexical data and/or knowledge bases. In order to achieve reliable semantic accuracy in the extraction process, we think the definition text should be parsed.

Previous works have described how semantic knowledge – specifically taxonomic information and other semantic relations – can be automatically extracted based on the regularity that occurs in definitions, both in their structure and in the recurring and systematic use of a limited set of *defining formulae*. The extraction of the 'genus' term takes advantage of the definition structure; the 'genus' term is usually, but not always,[1] the syntactic head of the defining phrase and the head(s) of the defining phrase can be identified with good results by means of heuristic procedures based on combinations of lexical categories to the left and right of the syntactic head (see Calzolari 1984; Chodorow et al. 1985). As for the semantic information contained in the 'differentia' part of the definition, its extraction is based on the observation that there are defining formulae in the definitions that systematically express conceptual categories, as well as semantic relations. Computationally, these defining formulae have been expressed as pattern-matching procedures which search for such occurrences of word forms as well as their co-occurrences within the defining phrases; this extraction procedure also yields promising results (see Markowitz et al., 1986; Calzolari & Picchi, 1988).

The Acquilex[2] project has adopted a two-stage strategy for extracting semantic knowledge from Italian dictionary definitions. During the first stage, a broad coverage Italian grammar provides an organized structure corresponding to an initial syntactic analysis

for each dictionary definition. There are two main reasons for parsing the definition. First, it is possible to abstract away from variations in the surface realization of the same pattern which exist regardless of the regularity typical in a dictionary. Second, the results are expected to be more reliable because we can specify a given level of embedding at which the defining formula is to be found, rather than accepting the defining formula no matter where it occurs, and because the real extraction process consists of identifying the relevant complements of the defining formulae and so accessing the structural information again yields more reliable results (see Montemagni & Vanderwende for a discussion of string patterns versus structural patterns.). During the second stage, a pattern-matching mechanism maps structural patterns onto the syntactic analysis computed at the previous stage, thereby deriving and making explicit the semantic knowledge implicitly stored in any standard printed dictionary. The general framework we are using, and tailoring for our purposes, was originally developed by Jensen and Binot for acquiring the semantic information necessary for the resolution of prepositional phrase attachment ambiguities (Jensen & Binot, 1987). Others have also accepted the use of syntactic analyses and structural patterns for some time now (Klavans, 1990, Ravin, 1990, and Vanderwende, 1990, all of which use the PLNLP English Parser to provide the structural information).

This paper focuses on the first stage of the extraction process, computing a syntactic analysis for each dictionary definition. This stage is crucial; it creates the data structures on which to operate during further processing stages and thereby determines the quantity and the quality of the information that can be extracted. We first experimented with the syntactic analyses as they are computed by a broad coverage Italian grammar. With this grammar it is possible to produce, on average, one parse per definition. In addition, by setting a switch, it is possible to force a parse of any desired category (NP, VP, etc.); and even if no parse is available for the entire string, pieces can be assembled, or "fitted," together so that there will always be some analysis for any given string. Although the output of the broad coverage grammar already was adequate, we chose to add a post-processor that would modify and improve the parses based on the peculiarities of the dictionary text type. The post-processor thus captures the differences between general text and dictionary text, a contrastive study that would not be possible if a dictionary-specific parser were constructed. We found that the post-processor is a very small component as compared to the broad-coverage grammar, reflecting our preliminary observations that the constructions found in dictionaries are as complex as, and not very different from, those of general text. Making use of a broad-coverage grammar, followed by a small post-processor, provides a robust parsing tool that is both efficient with respect to the reusability of components and interesting for contrastive reasons.

2. Syntactic parsing

The broad-coverage Italian grammar that was used for this study was written following the general strategy called the 'relaxed approach' aimed at accepting unrestricted input text (Jensen, 1986, 1988, 1989). Sentences are analyzed according to syntactic information formalized in augmented phrase structure rules with a bottom-up, parallel parsing algorithm, producing an attribute-value analysis structure that can be displayed as a parse

tree (Heidorn, 1975). The lexicon which supports this analysis contains very limited information (parts of speech, morphology, and essential word class features). A grammar constructed in this way computes preliminary syntactic sketches that are syntactically consistent, but not necessarily semantically valid. The analyses contain syntactic and – whenever possible – functional information, but no semantic or other information beyond the functional level. In Italian, in some cases, even the functional roles cannot be assigned on the basis of purely syntactic information but only after background (semantic and/or contextual) information has been acquired and evaluated within the initial analysis.

The analysis of a sentence using only syntactic information may contain many ambiguities. We just mentioned the ambiguity of assigning functional roles. Attaching modifiers to their appropriate heads is the other main source of ambiguity. The strategy adopted for dealing with both kinds of ambiguity is that of packing the different syntactic descriptions into the same structure whenever possible. For attachment ambiguity, the solution is to attach modifiers to the closest possible head, and to mark alternative attachment sites so that they can be tracked down for later semantic processing. For functional ambiguity, we code the possible interpretations within the same structure in order to have them ready for further processing stages. This is the reason why we usually think of the resulting analysis as a 'syntactic sketch'. This attachment and assignment strategy, which allows the grammar to produce on average one parse per sentence, eliminates any combinatorial explosion while preserving all the necessary information.

3. Parsing dictionary definitions with a broad-coverage Italian grammar

The first question to be answered at this point is whether and how well dictionary definitions can be analyzed by a general purpose grammar. The formulaic language of dictionary text mentioned above reflects the frequent occurrence of lexical and syntactic patterns expressing particular conceptual categories or semantic relations, and the higher frequency of defining generic terms (see Calzolari, 1984). These formulae, however crucial to the extraction of semantic information, can be considered almost irrelevant from the point of view of parsing because the variety of syntactic constructions in which these formulae are manifested can be compared to that of text corpora. And this is also true with respect to the vocabulary used within definitions since, unfortunately, none of the Italian dictionaries uses a restricted vocabulary (unlike the LONGMAN DICTIONARY OF CONTEMPORARY ENGLISH). These two factors combined make a robust analysis even more necessary. Such a variety of lexical choices and phrasal constructions in dictionary definitions poses, therefore, the same range of problems a parser is faced with in analyzing ordinary texts.

Dictionary text does of course differ from general text in some predictable ways. First, and most obvious, the definition text rarely forms a complete sentence. Fortunately, the syntactic form of the definition text is largely predictable from the part of speech of the definiendum. It is therefore important that the parser provide a switch indicating whether the input should be parsed as a nominal, verbal, adjectival, adverbial, or prepositional phrase or as a relative clause, depending on the part of speech of the definiendum and on the definition itself.

For example, the parse trees[3] in Figure 1 show the parse of a very simple definition in Garzanti for the noun "arancia" (orange) before and after the switch has been set which forces an NP analysis. The definition reads: "frutto dell'arancio" (fruit of the orange tree).

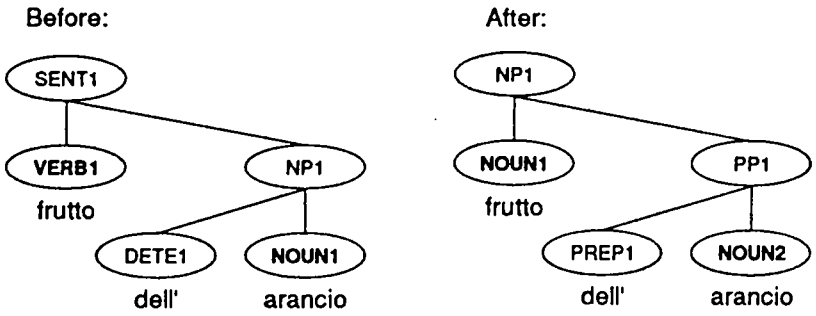


Figure 1. Parse trees for the definition text "frutto dell' arancio".

In the first parse, without an NP, the definition has been analyzed as a complete sentence with an empty subject, headed by the verb "fruttare" (to yield, in the financial field), followed by the noun phrase "dell'arancio" as object. According to this analysis, the string would be translated as "I yield some orange tree". While this analysis is syntactically valid, it is not semantically well-formed; this interpretation can be ruled out only on the basis of semantic information. First, it is very unlikely (if not impossible) for the noun "arancio" to be the object of the verb "fruttare". Second, the partitive determiner "dello" cannot premodify a countable singular noun.[4] The more appropriate second parse is obtained by forcing the analysis of the input string to be an NP. Thus, although a sentential parse is possible, the correct NP parse is computed given that this text is the definition text of a noun definiendum. The category switch is an essential tool because it allows the correct phrasal parse to be computed without having made any modification to the broad-coverage grammar.

Second, definition text also differs from general text because it does not always form even a complete phrase but often only fragments of phrases (e.g. obligatorily transitive verbs without objects). It is therefore necessary that a parser provide a form of 'fitted parsing' (Jensen et al. 1983) for handling fragments and for handling gaps in the grammar itself to ensure that the parser never fails to produce an analysis. 'Fitted parsing' is accomplished by a set of procedures which assign a reasonable approximate structure to the input in cases where no parse covering the entire string could be computed. Such a rough parse is still useful as input for further processing stages and for the extraction procedure itself (even if the results of this extraction have to be treated differently from those derived from a complete analysis). Examples of the results of the fitting procedure applied to dictionary definitions will be given in the following section.

Using only the broad-coverage Italian grammar and the parser described above, we began parsing definition text extracted from II NUOVO DIZIONARIO GARZANTI and the

ITALIAN DMI DATABASE, mainly based on the Zingarelli dictionary. It didn't take a lot for us to identify two main areas of the grammar which needed to be tailored in order to give more appropriate parsing results for dictionary text.

1. Resolution of ambiguous assignment.

The default strategy for attachment ambiguity, namely attachment to the closest possible head, should sometimes be changed for dictionary text. In this way, some attachments which would remain ambiguous in ordinary texts can be disambiguated in the context of dictionary definitions. This is the case, for instance, with the attachment of post- modifiers to coordinated genus terms of certain classes. Similarly, the functional role assignment, ambiguous in general text, almost always can be disambiguated in the context of dictionary definitions. We assume that constructions used within dictionary definitions are always in unmarked SVO order and that the ambiguity stemming from potentially marked ordering of sentence constituents (such as SOV, OVS, and so forth) is very unlikely to occur in this specific context.[5]

2. Analysis of specific dictionary-language constructions.

Definition texts should be seen as fragments of wider text corpora. Very rarely do they appear as complete sentences, in which case they are exceptions within the definition language. They are usually formulated as NP, VP, AdjP, AdvP, PP, or relative clauses; but because they are condensed fragments of real texts, obligatory elements are sometimes elided, which makes the definition syntactically ill-formed and interpretable only by reference to a wider context. From this perspective, it is often the case that syntactic deviance from the point of view of a general grammar is a typical occurrence within dictionary definitions. Such is the case with noun definitions formulated as a noun phrase premodified by a prepositional phrase, where the PP specifies the usage domain of the word sense expressed by the NP. Because a PP-NP construction (with the PP pre-modifying the NP) is a syntactically deviant order within the core grammar of Italian, the grammar is unable to produce an NP node covering the whole input string, in spite of the switch forcing an NP analysis.

These observations necessitate a revision of the grammar output in order to make the extraction of semantic information from natural language definitions more efficient and reliable. This revision has been carried out (a) by ruling out some ambiguous constructions, and (b) by handling and regularizing otherwise ill-formed input. The next section will describe when and how these tasks are performed in relation to the whole parsing procedure.

4. Disambiguating and reshaping the syntactic analysis of the definitions

We decided not to intervene in the general grammar itself, which should remain restricted, in our opinion, to the description of the central, agreed-upon grammatical structures of language. The choice was made to revise the initial syntactic analysis during a post-processing stage. The disambiguation task is carried out by a module specifically conceived for this purpose, the Dictionary Definition Disambiguator (DDD). This component, still in an embryonic stage, has been designed to resolve, whenever possible, what was left undecided during the first stage of processing. The task of reshaping incomplete parses is handled by modifying the fitting procedure to deal properly with the ill-formed but, in the context of dictionary language, common constructions. This minor addition to the overall architecture of the general parsing system led to a marked improvement of the parsing results. Since these parses are the input to the component

for extracting semantic information from dictionary definitions, there was also a marked improvement in the quality and reliability of the semantic information thus extracted. Let's illustrate the way the analysis produced by the general grammar is revised (disambiguated or reshaped) during this post-syntactic stage.

Disambiguation is concerned with attachment as well as assignment problems. The example in Figure 2 shows the resolution of prepositional phrase attachment to coordinated genus terms such as "atto" (act), "effetto" (effect), "processo" (process). This pattern is typical of the definition of deverbal nouns; the PP which follows the (conjoined) genus terms indicates the verb from which the definiendum is derived. In the Garzanti definition for "computazione" (computation): "atto, effetto del computare" (act, effect of computing), the genus terms are "atto" and "effetto":

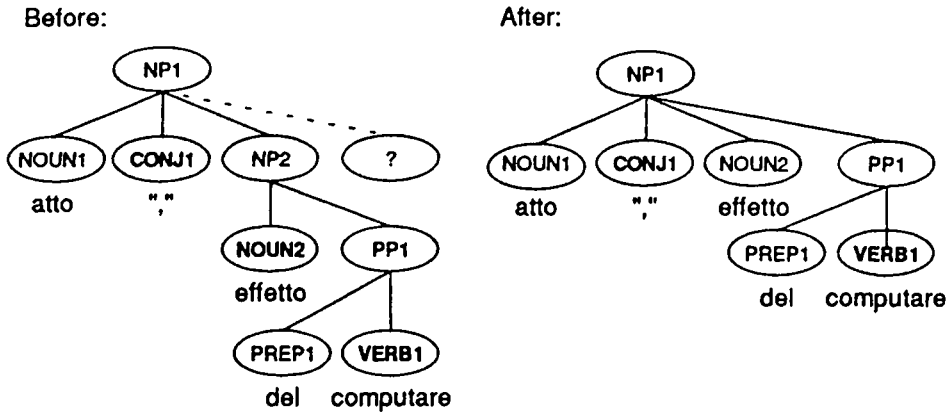
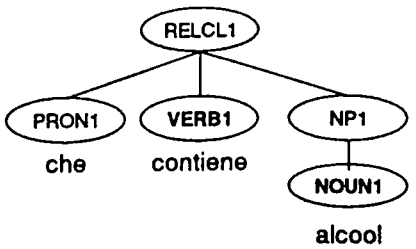


Figure 2. The reattachment of prepositional phrases in DDD

In the "before" parse, the core grammar has applied the default attachment strategy. PP1 ("del computare") is attached to the closest available head, "effetto", and the alternative attachment site, the nominal phrase covering the coordinated genus terms, is indicated by a question mark. This parse would only allow the semantic relation "effetto del computare" to be extracted from the parse. The "after" parse shows the analysis after the revision performed by the DDD. After checking the syntactic and lexical conditions which hold for this type of disambiguation, the analysis is refined by reattaching PP1 to the nominal phrase covering the coordinated genus terms, so that now the PP modifies both of the genus terms and not only "effetto". This analysis allows the semantic relation "atto del computare" as well as "effetto del computare" to be extracted for the noun "computazione", thus increasing the reliability and completeness of the semantic information extracted.

A second example, also handled by the DDD, illustrates the resolution (sometimes only a reduction in the ambiguity range) of ambiguous functional roles. Relative clauses are the only context in dictionary language where different word orders are equivalent from the point of view of markedness (that is, are unmarked). Of the four possible orders within relative clauses - SVO, SOV on the one hand and OVS, OSV on the other - only the

SOV configuration appears to be marked. Functional role ambiguity therefore occurs only in relative clauses for which agreement in number and gender between subject and verb cannot determine which NP is the subject because both candidate NPs – or the only existing one – agree with the verb. Consider the Garzanti definition, in the form of a relative clause, for the adjective "alcolico" (alcoholic): "che contiene alcool" (which contains alcohol). In general text, it is not clear which NP is subject and which is object; the clause might be paraphrased in English either as "it contains alcohol" or as "alcohol contains it". Since both candidate NPs agree with the verb, subject-verb agreement cannot serve to resolve the ambiguity. Figure 3 shows the syntactic parse tree followed by a simplified attribute-value record structure that indicates functional information. The attribute SUSPNPS (suspicious NPs) points to NP1, "alcool", which at this first stage of analysis is interpreted as a candidate for both functional roles, SUBJECT and OBJECT:



SEGTYPE	RELCL	
STR	che contiene alcool	
PRMODS	che	
HEAD	VERB1	
PSMODS	NP1	alcool
SUSPNPS	NP1	alcool

Figure 3. Ambiguous functional roles in a relative clause.

The DDD can resolve this ambiguity completely on condition that the definition is formulated as a complete relative clause, because there is an implicit convention that top level relative clauses have SVO order. The DDD automatically assigns the suspicious NP1 ("alcool") as an OBJECT in the revised record structure and now the appropriate semantic information can be extracted reliably.

SEGTYPE	RELCL	
STR	che contiene alcool	
PRMODS	che	
HEAD	VERB1	
PSMODS	NP1	alcool
OBJECT	NP1	alcool

Figure 4. Disambiguated record structure in Figure 3.

Many adjective definitions are syntactically formulated as relative clauses. For all of them, the initial ambiguity detected by the general grammar can be resolved easily and with certainty as described above. However, for relative clauses embedded within definition texts, the functional role ambiguity cannot be resolved so easily or with the same degree of certainty. Although SVO, OVS, and OSV are all unmarked orders in the context of relative clauses, SVO and OSV configurations are the preferred ones, while the sequence OVS seems to occur only when the functional role assignment can be determined on the basis of the agreement. By reducing the possible configurations of embedded relative clauses to only the unmarked ones, and by restricting the occurrence of the OVS order to cases in which the assignment is defined on the basis of the agreement, the dictionary language seems to limit functional ambiguity. Yet, there still remain some cases of functional ambiguity to be resolved by higher semantic or contextual information, just as in general text. It is therefore reasonable to claim that while word order unmarkedness is a matter of strong preference, it cannot possibly support a watertight resolution strategy.

All the previous examples are handled by the DDD. The input to this module is an analysis provided by the general grammar that is complete and needs only to be refined with respect to the ambiguity in attachment or functional role assignment of the initial syntactic analysis. This module is in charge of selecting the best interpretation from which to extract semantic information based on conventions discovered in dictionary definition language. These conventions are easily captured as conditions on the initial analyses; and though the number of these conditions is actually very small, their combined effect serves to greatly improve the quality of the dictionary parses and also of the semantic information extracted from them.

A final example illustrates the case where the general grammar does not, and should not, produce a complete analysis. Such cases are handled appropriately by the fitting procedure which we have tailored to produce the correct results for the dictionary definition domain. Consider the analysis of the Garzanti definition of "nettare" (nectar), defined as "nella mitologia classica, la bevanda degli dei" (within classical mythology, the drink of the gods). As with many definitions, the domain in which this specific sense of the noun holds (namely, classical mythology) is expressed as a PP preceding the actual noun definition. This PP - NP sequence is not an acceptable nominal phrase in general Italian (i.e. in text corpora) and so the "before" parse has a top level node XXXX indicating that there is no single constituent that covers the entire string.

In order to extract semantic information reliably, we have tailored the fitting procedure to rebuild the initial syntactic analysis given that the string occurs in the context of dictionary definitions. All that was required was to allow the fitting procedure to build an NP from an NP premodified by a PP in a definition text and then automatically the fitting procedure has available a top level NP node with the genus term "bevanda" (drink) as its head. By rebuilding the initial syntactic analysis in this way, semantic information can now be extracted easily and reliably. With this strategy, we do not need to modify the general purpose grammar to handle what is ill-formed outside the context of dictionary definitions. As it is, the distinction between ill- and well-formed input is not always clear, even in regular text, and so the fitting procedure allows what appears to be ill-formed in the general text to become regular with respect to some specialized use.

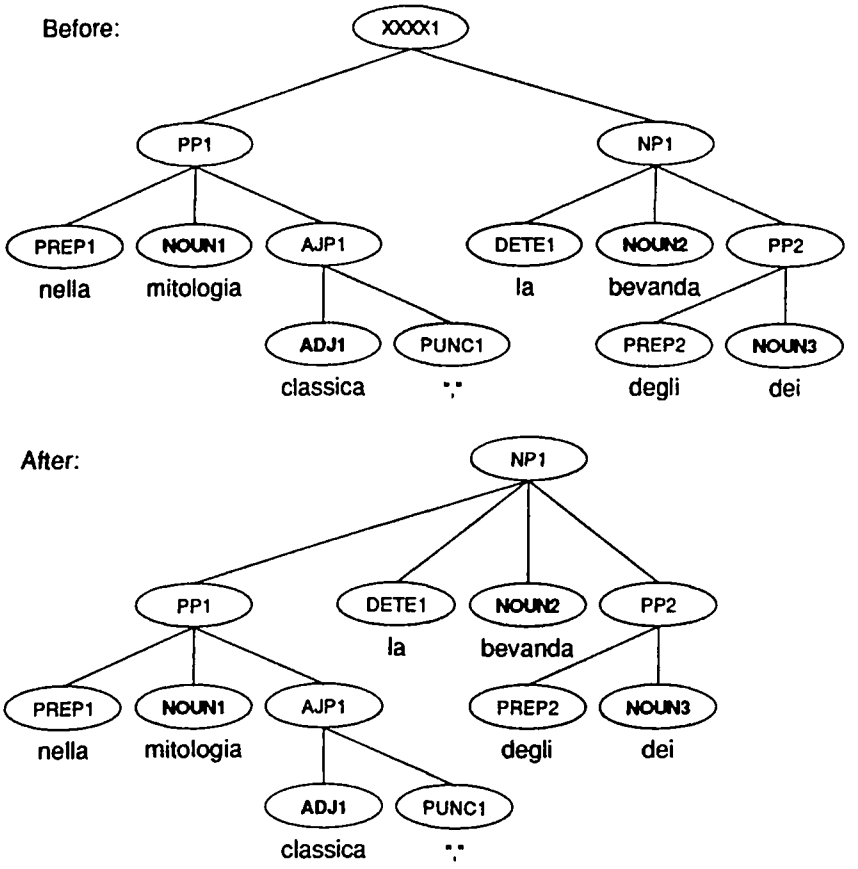


Figure 5. Reshaping a fitted parse: NP with pre-modifying PP.

5. Concluding remarks

The research described here is still in progress. The work carried out so far suggests that dictionary language differs from the language of general text only along a few, easily defined, parameters; certainly it cannot be defined as a specialized language given that it does not operate in a specialized domain. This contrastive work was possible after parsing a significant subset of definitions with a general purpose grammar of Italian. Dictionary language differs from the language in general text because it has less ambiguity in the attachment and assignment of constituents and because text that is considered ill-formed outside the context of the dictionary is allowed. If we can reliably parse dictionary language, then those analyses can in turn be used to extract semantic information accurately.

Instead of modifying the core Italian grammar to handle dictionary definitions (modifications that would be ad hoc with respect to the coverage of the general grammar), we decided to integrate the parser with a post-processing component that operates on the

initial syntactic component. This post-processor is in fact a very small component, commensurate with the very small number of differences between general and dictionary text. Its tasks are to rule out the ambiguity in the initial syntactic analysis and to reshape the initial analysis where the general grammar cannot construct a phrase covering the definition string, i.e. in cases where the string is ill-formed in contexts outside the dictionary.

As more definition sets are parsed, the post-processor can be incrementally revised and improved while leaving the core grammar intact and reusable for other purposes. This separation allows us to catalogue the actual differences between dictionary and general text. In the end, we hope to arrive at a system that provides the best structured information from which to extract semantic information, and also to have a detailed description of the language used in dictionaries.

Endnotes

1 The "not always" cases are limited and predictable; for example, semantically empty hypernyms such as "gruppo" (group), "insieme" (set), "parte" (part), "porzione" (portion), which express non-taxonomic information.

2 The Acquisition of Lexical Knowledge from Machine Readable Dictionaries, Esprit BRA 3030.

3 The parse trees in this paper are altered representations of actual machine output, which IBM ASD has withheld from publication. Here, heads of constituents are directly below their parent node and the nodename is in bold.

4 Mass/count information is so complex that it must be treated semantically

5 In Italian, given a simple transitive sentence, with a verb, a subject, and an object, all the six permutations of the three constituents are grammatically acceptable. SVO is the basic, dominant order and variations on this basic word order are generally marked because they produce semantic and pragmatic effects. The default word order varies according to whether it appears at the main clause level or in relative or interrogative clauses. This implies that the assignment of functional roles cannot be based, as happens for instance in English, on the structural configuration of the sentence. Ambiguous subject/object assignment cases occur when there are two NPs (or just one, given that Italian is also a pro-drop language) that agree with a transitive verb. In this case, it is impossible to assign the functional roles without semantic and/or contextual information about the predicate and its arguments.

Bibliography

CALZOLARI, N. (1984): "Detecting Patterns in a lexical Data Base", in Proceedings of the 10th International Conference on Computational Linguistics, Stanford (CA), 170-173.

CALZOLARI, N., PICCHI E. (1988): "Acquisition of Semantic Information from an On-Line Dictionary", Proceedings of the 12th International Conference on Computational Linguistics, Budapest, 1988, 87-92.

CHODOROW, M. S., BYRD R. J., HEIDORN G. E. (1985): "Extracting semantic Hierarchies from a large on-line dictionary", Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics, University of Chicago, Chicago, 8-12 July, 299-304.

HEIDORN, G. E. (1975): "Augmented Phrase Structure Grammars". in: Schank and Nash-Webber, eds. Theoretical Issues in Natural Language Processing. Association for Computational Linguistics.

JENSEN, K. (1986): "PEG 1986: A Broad-coverage Computational Syntax of English", Unpublished paper.

JENSEN, K. (1988): "Issues in Parsing", Proceedings of the Symposium on Natural Language at the Computer, published by Springer Verlag.

JENSEN, K. (1989): "A Broad-coverage Natural Language Analysis System", Proceedings of the International Workshop on Parsing Technologies, Carnegie Mellon University, 28-31 August 1989.

JENSEN, K., BINOT J. L. (1987): "Disambiguating Prepositional Phrase Attachments by Using On-Line Dictionary Definitions", Computational Linguistics, 13, 3-4.

JENSEN, K., HEIDORN G.E., MILLER L.A. , RAVIN Y. (1983): "Parse Fitting and Prose Fixing: Getting a Hold on Ill-formedness", in the American Journal of Computational Linguistics, 9, 3-4.

KLAVANS, J., CHODOROW M.S., WACHOLDER N. (1990): "From Dictionary to Knowledge Base via Taxonomy" in Electronic Text Research, University of Waterloo, Centre for the New OED and Text Research, Waterloo, Canada.

MARKOWITZ, J., AHLSWEDE T., EVANS M. (1986): "Semantically significant Patterns in Dictionary Definitions", in: Biermann A. (ed.) Proceedings of the Association for Computational Linguistics (ACL) 24th Annual Meeting; 10-13 June; New York, 112-119.

MONTEMAGNI, S., VANDERWENDE L. (1992): "Structural Patterns vs. String Patterns for Extracting Semantic Information from Dictionaries", to appear in: Proceedings of COLING-92, 20- 28 July; Nantes, France.

RAVIN, Y. (1990): "Disambiguating and Interpreting Verb Definitions" in Proceedings of the Association for Computational Linguistics (ACL) 28th Annual Conference,

VANDERWENDE, L. (1990): "Using an on-line Dictionary to disambiguate Verbal Phrase Attachment", in: Proceedings of the 2nd IBM ITL Conference on NLP, La Defense, Paris, 13-15 March.

KEYWORDS: computational lexicography, on-line dictionaries, specialized grammar, parsing dictionaries, semantic knowledge extraction, reusability of grammars.