

# Building a Treebank for Italian: a Data-driven Annotation Schema

Cristina Bosco\*, Vincenzo Lombardo<sup>†</sup>,  
Daniela Vassallo\*, Leonardo Lesmo\*

\* Dipartimento di Informatica, Università di Torino,  
c.so Svizzera 185, 10149, Torino (Italy)

<sup>†</sup>DISTA, Università del Piemonte Orientale "A. Avogadro",  
c.so Borsalino 54, I - 15100 Alessandria (Italy)  
{bosco, vincenzo, vassallo, lesmo}@di.unito.it

## Abstract

Many natural language researchers are currently turning their attention to treebank development and trying to achieve accuracy and corpus data coverage in their representation formats. This paper presents a data-driven annotation schema developed for an Italian treebank ensuring data coverage and consistency between annotation of linguistic phenomena. The schema is a dependency-based format centered upon the notion of predicate-argument structure augmented with traces to represent discontinuous constituents. The treebank development involves an annotation process performed by a human annotator helped by an interactive parsing tool that builds incrementally syntactic representation of the sentence. To increase the syntactic knowledge of this parser, a specific data-driven strategy has been applied. We describe the cyclical development of the annotation schema highlighting the richness and flexibility of the format, and we present some representational issues.

## 1. Introduction

The relevance of linguistic corpora to several areas of natural language processing, is bringing about a worldwide development of such a resource in many languages. And the success of corpus-based methods can cross-linguistically increase with the availability of syntactic annotated corpora known as *treebanks* (see, e.g., (ATALA, 1999) for report of treebanks for German, Spanish, Chinese, Polish).

In general, syntactic annotation schemata of existing treebanks try to achieve a trade-off between representation accuracy and corpus data coverage satisfying some requirements of theory-independence (Skut et al., 1997) and annotation consistency between the analyses of different phenomena (Marcus et al., 1994). Data coverage requires a formalism that is able to represent all the types of specific linguistic phenomena occurring in the corpus, catching most of the peculiarities of the represented natural language.

In this paper, we present the main issues of an Italian treebank project, the Turin University Treebank (TUT), currently in a preliminary development phase. We describe a data-driven cyclical development process of the treebank, ensuring both data coverage and consistency between annotation of related phenomena. In fact, we define a core representation format and augment it to yield a representation for each sentence of the corpus; consistency between structures representing similar phenomena is tested.

The paper is organized as follows. In the next section we describe in detail the annotation schema, and discuss its motivations comparing our experience with other works in this area. The third section present the TUT development cycle and the current status of the project. Then in the last section we show representations of some relevant linguistic phenomena found in the corpus.

## 2. The annotation schema

In general, a treebank annotation schema should point out the relevant aspects of syntactic constructions, and should represent in an appropriate and distinct manner as many as possible linguistic phenomena. The choice of a specific format for the representation of the language structure should depend on the capability of representing the phenomena of interest with the sort of representational elements that the format provides. The Penn Treebank implements a syntactic annotation schema based on phrase structures, and provides some non-context free annotational mechanisms to represent discontinuous constituents (Marcus et al., 1994); the Prague Dependency Treebank has a dependency-based representation naturally oriented to predicate-argument structure (Bemova et al., 1999); the NEGRA treebank adopts a mixed framework combining phrase-structure and dependency grammar representation formats (Skut et al., 1999).

A phrase-structure formalism postulates the existence of non-terminal syntactic categories, and includes in parse trees both nodes labeled with terminal symbols (words) and nodes labeled with non-terminal symbols (phrases). By contrast, a formalism based on dependency grammar (DG) does not include non-terminal symbols and describes the syntactic structure in terms of binary relations on pairs of words, a head and a dependent respectively, forming a dependency tree. Dependency-based representation formats are more proximate to semantics and allow for an immediate mapping of trees onto predicate-argument structures.

Further differences in treebanks schemata are also related to language dependent aspects. For example, in the Penn Treebank, traces are used in the representation of a variety of phenomena involving long distance dependencies (such as topicalization and wh-movements); similar constructions in the NEGRA treebank involve the presence of crossing edges in syntactic trees. The two different approaches are motivated by the low frequency of discontinuous constituents in English, and the respectively high frequency in German; in representing

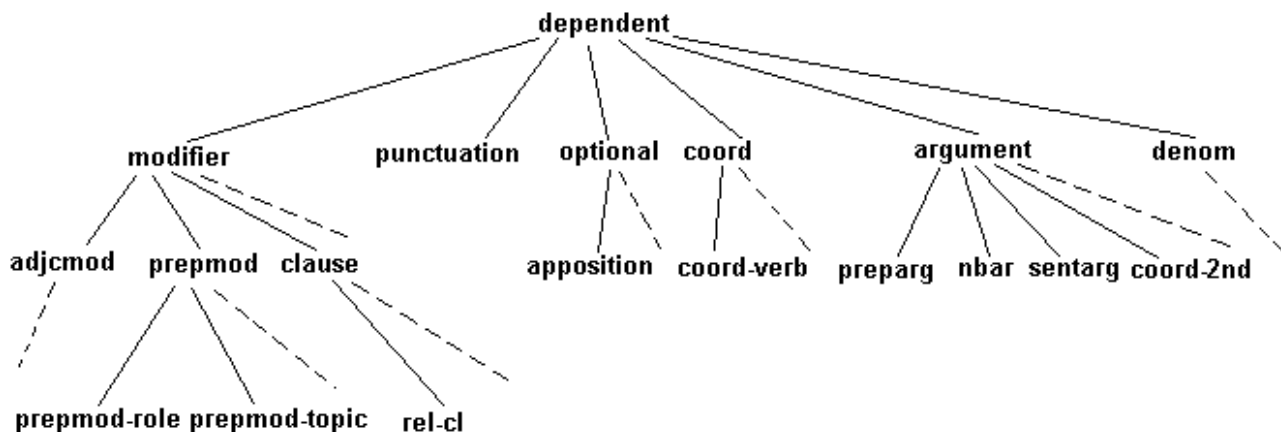


Figure 1. The upper levels of the dependency relations hierarchy.

German, an annotation format with traces would bring about a great proliferation of co-references (Skut et al., 1997).

In the rest of this section we describe the TUT annotation schema.

### 2.1 The Turin University Treebank schema

From the reported experiences, dependency-based formats are specially adequate for non-configurational languages, for example free-word order languages, where the role of syntactic units does not depend strictly on word order (cfr. Skut et al., 1997). Since Italian is characterized by a partial configurationality at the verbal level, we have adopted a dependency-based formalism centered upon the notion of predicate-argument structure.

To represent discontinuous constituents and phenomena found in the corpus, such as null subject and coordination, the core paradigm has been augmented with a trace-filler notation. The rare usage of discontinuity in Italian allows for the adoption of this representational strategy (like in Penn Treebank) for non-local dependencies, thus allowing for a clear distinction with respect to local dependencies.

The TUT annotation schema pursues descriptive richness and representational flexibility. Richness is achieved through a data-driven refinement of basic

representation. Flexibility results from a hierarchical organization of dependency relations from generic to specific ones (see Fig. 1). Upper levels of this taxonomy represent the basic kind of dependencies, such as modifier and argument; lower levels show the specific relation involving specific grammar classes, for example a preposition (prepositional modifier, *prepmo*; prepositional argument, *preparg*).

If the annotator cannot decide for a unique very detailed grammatical relation he/she can select an underspecified relation from upper levels of the hierarchy. For example, the relation between a noun (head) and an adjective (dependent) can be labeled with adjectival modification (*adjcmod*) and further refined on the basis of the adjective features (*adjcmod-qualif*, *adjcmod-ord* ....) (see Fig. 2).

A similar problem has been treated by the Penn Treebank researchers in a similar way, in the phrase-structure format of the treebank. When an annotator is sure that a sequence of words is a major constituent but is unsure of its syntactic category, he/she can use the more general X constituent label (Marcus et al., 1993). This approach features a binary level of underspecification; in our schema, the annotator can achieve smoother the refinements on several levels of grammatical relations.

### 3. Turin University Treebank development cycle

The construction of a treebank is a particularly labor-intensive and time-consuming task usually performed by human annotators with the help of software tools. Usually the annotation process occurs in two phases: a fully automated Part Of Speech (POS) tagging and a syntactic annotation that can be performed in different ways. For example, in the Penn Treebank (Marcus et al., 1993) syntactic bracketing consists of the manual correction of an automated parsing output and annotator possibly "glues" together disconnected syntactic chunks. In the NEGRA treebank (Brants et al., 1999) the annotation is an interactive process that the annotator can stop at any point to correct or alter the structure automatically generated. The development of the TUT follows the approach of the NEGRA project in using an interactive parser. The major difference consists in the parser

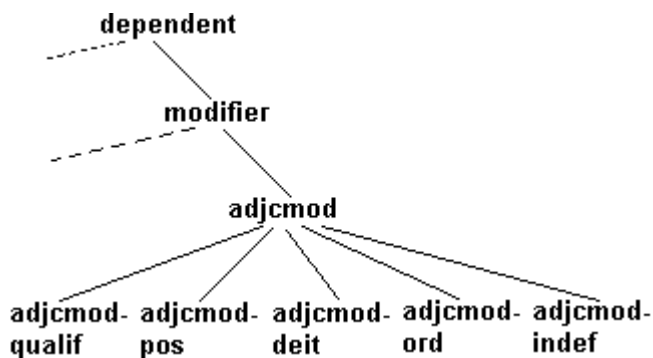


Fig.2: The part of the hierarchy related to adjectival modifiers.

dependency relations operated during corpus annotation and motivated by the need of a semantics-oriented

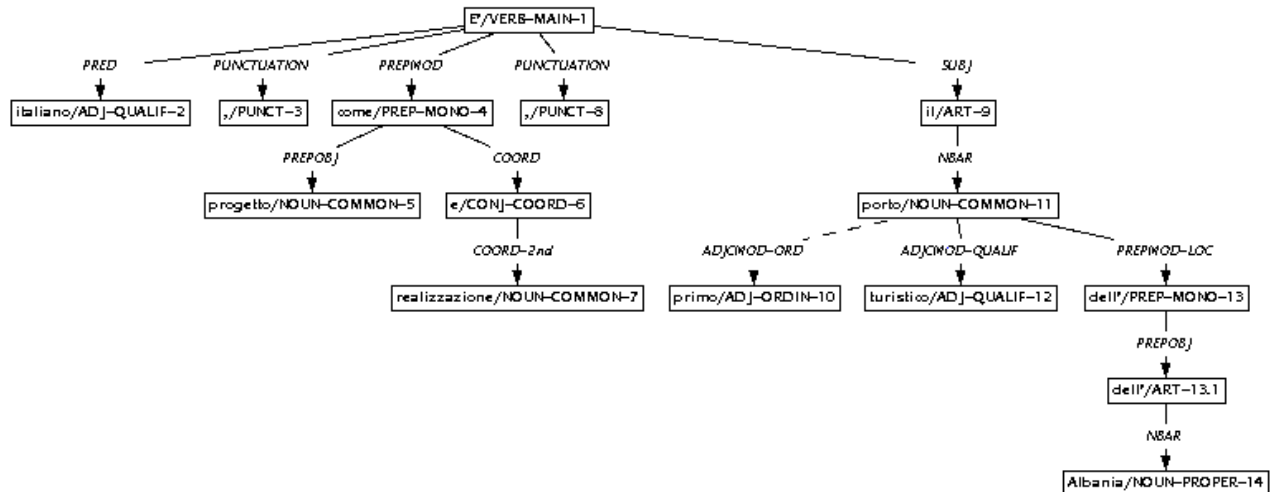


Figure 3. The syntactically annotated tree representing, in the daVinci graphic format, the sentence "E' italiano, come progetto e realizzazione, il primo porto turistico dell'Albania." (The first tourist port of Albania is Italian as project and realization). Dotted lines represent left dependencies.

strategy, which, in the case of TUT, pursues a psycholinguistic approach.

In the rest of this section we present the annotation process and the parsing tool implemented for the TUT; then we describe the data-driven refinement of the annotation schema.

### 3.1 The annotation process

The corpus of texts we are working on in the treebank project is a selection of articles from Italian newspapers about Albania. The corpus has been collected by psychology researchers to analyze how the view of Albania and Albanians has changed over recent years in Italy.

We have annotated a thousand sentences automatically through an interactive parsing tool specially designed for this purpose. An annotator interacts with the parser on POS tagged sentences (the tagger is described in (Boella & Lesmo, 1998)). The graphical interface daVinci (see <http://www.tzi.de/~davinci/>) displays the structure on the screen for the annotator's decision (see example in Fig. 3). The syntactic representation is built incrementally: for each word from left to right, the parser, on the basis of the current grammar, incorporates the current input word in a partial, but fully connected, tree that the user can accept or reject. If the annotator accepts the proposed structure, the parser continues the processing with the next input word, otherwise the parser suggests alternative structures. Traces are possibly inserted manually during the process. The grammar consists of individual relations (rules) on pairs of word categories.

Fig. 4 shows the rules that license the syntactic tree presented in Fig. 3. For example, the rule 6 states that the PREPOBJ relation can label the dependency between a PREP-MONO (monosyllabic preposition) and a NOUN-COMMON, where the noun follows the preposition (Right) at any linear position with respect to other dependents on the right (Any); the rule also states that the preobj relation is a necessary dependent of the preposition (minimum cardinality 1, maximum cardinality

1), with a constraint on the co-occurrence of dependency relations ( $\{<p.4>\}$ ). This constraint states that this relation is incompatible with other possible relations with which is not associated the same co-occurrence index, for example with the following :

**(PREPOBJ PREP-MONO PRON-INDEF Right.Any 1.1 {<p.15>}).**

In the treebank development, these grammar rules are viewed as basic "bricks" through which dependency structures are incrementally generated to represent corpus

1. (PRED VERB-MAIN ADJ-QUALIF Any.Any 1.1 {})
2. (PUNCTUATION X PUNCT Right.Any 0.Any {})
3. (PREPMOD VERB-MAIN PREP-MONO Any.Any 0.Any {})
4. (COORD X CONJ-COORD Right.Any 0.1 {})
5. (COORD-2nd CONJ-COORD X Right.Any 1.1 {})
6. (PREPOBJ PREP-MONO NOUN-COMMON Right.Any 1.1 {<p.4>})
7. (SUBJ VERB-MAIN ART Any.Any 0.1 {<s.3>})
8. (NSAR ART NOUN-COMMON Right.Any 1.1 {<n.1>})
9. (ADJCMOD-ORD NOUN-COMMON ADJ-ORDIN Left.Any 0.1 {})
10. (ADJCMOD-QUALIF NOUN-COMMON ADJ-QUALIF Any.Any 0.Any {})

Figure 4. The set of grammar rules that license the syntactic tree of Fig. 3

sentences.

To increase the syntactic knowledge of the parser, we have implemented the following data-driven strategy: during the treebank development, complex syntactic constructions (involving the combination of various rules) are stored in a database. The parser can then hypothesize and suggest to the annotator prebuilt structures without recomputing them from the basic bricks.

A further development of this annotation tool will include statistically driven methods to improve the reliability and

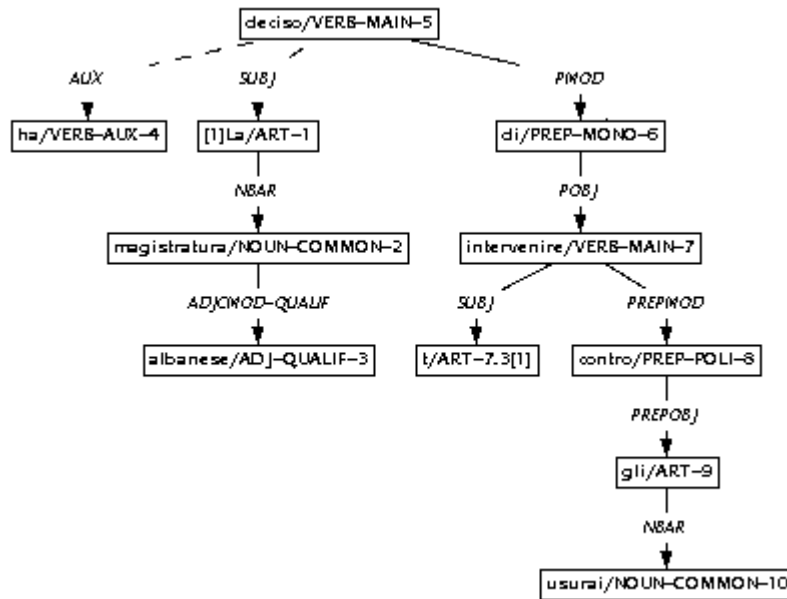


Figure 5. The dependency tree of the sentence "La magistratura albanese ha deciso di intervenire contro gli usurari." (The Albanian magistrature has decided to intervene against the usurers)

speed of annotation. Statistical data will be collected on the prebuilt structures.

### 3.2 The cyclical development of the annotation schema

Consistency in annotation of similar phenomena and easiness of further processing of treebank go together: for example, in the Penn Treebank project, the lacking of consistency was one of the main motivations for the development of a new annotation schema for the second release (Marcus et al., 1994).

The cyclical development of TUT representation format should ensure both data coverage and consistency between annotation of related phenomena. The extension of the schema during the training phase of the annotation process ensures the satisfaction of data-drivenness and coverage requirements; specific representational structures or more refined syntactic relations have been introduced in the schema to treat constructions and phenomena found in the corpus or to enrich in some form the representation. A cyclical comparison of implemented structures has help us to maintain the consistency between them. For this purpose we are compiling a detailed style-book in which the more strictly representational issues are collected and motivated. In two ways this compilation can be very useful: in helping annotators and in maintaining annotation consistency.

## 4. Representation of linguistic phenomena

In this section we review a few relevant phenomena encountered in the preliminary phases of development of the project.

The sharp distinction between head and dependents stipulated by dependency syntax, though useful in the

definition of semantic interpretation procedures, may cause difficulties in practice: in particular, for all kinds of constructions without a clear syntactic head such as ellipses like some forms of balanced structures (coordinations, comparatives i.e.). Other problems may come from phenomena such as null subject of infinitive complement clauses or implicit subject (pro-drop phenomenon), because of the surface alterations of predicate-argument structure (in terms of subcategorization frames) (see Lesmo & Lombardo, this volume).

### 4.1 Null elements

Including null elements in syntactic representation can be useful in many cases to represent predicate-argument structure long distance dependencies. For example, in the annotation tagset proposed for English in the Penn Treebank, null elements are used in representation of linguistic phenomena such as WH-movements, passive clauses, topicalization (or, in general, structures where constituents do not appear in their default position) and ellipsed materials (Marcus et al. 1994). Moreover, a null element co-indexed with the controlling NP is used to indicate which lexical NP is to be interpreted as the null subject of an infinitive complement clause.

In our formalism trees representing infinitive clauses contain empty subjects (trace); the null element is co-indexed via a numeric identifier with the subject of the tensed verb on which the un-tensed (infinitive) depends. In Fig. 5 an example of infinitive clause is showed: the subject of the dependent (infinitive) clause is represented by a trace (t / ART - 7.3 [1]) co-indexed with the head of the subject of the main clause ([1] La/ ART - 1); [1] is the numeric co-reference identifier of this co-indexing relation.

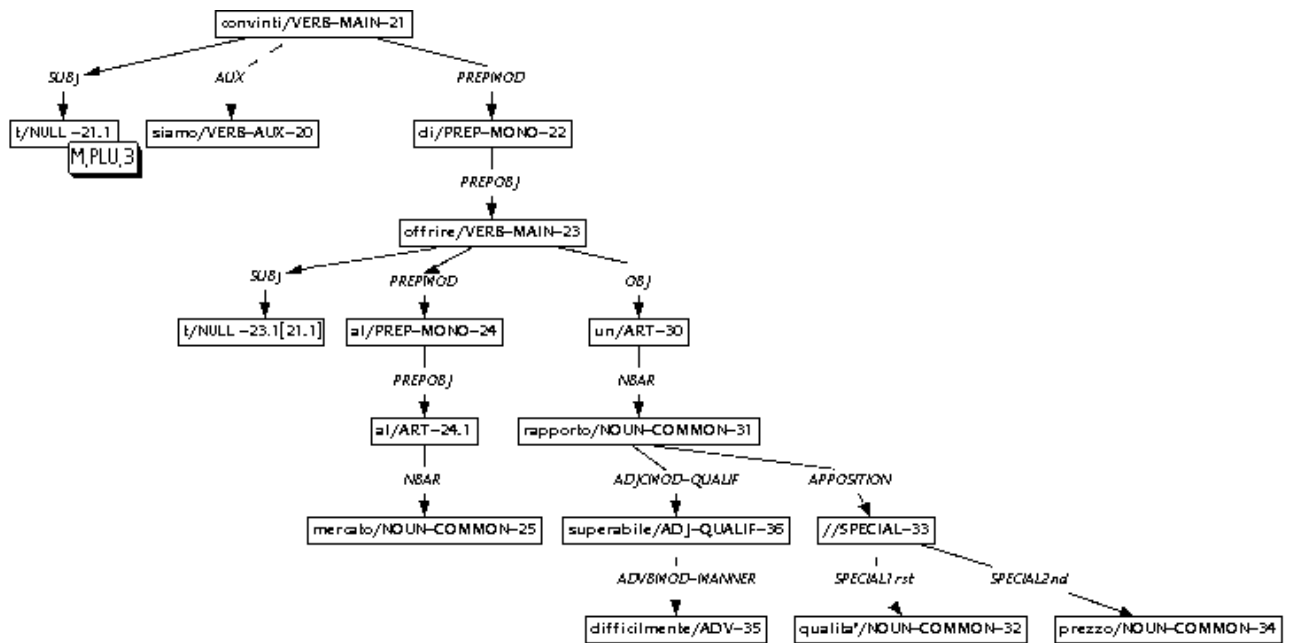


Figure 6 "...siamo convinti di offrire al mercato un rapporto qualità/prezzo difficilmente superabile..." (...[we] are convinced to offer to the market a proportion between quality and price difficult to exceed)

In Italian, as well as in most Romance languages, null subjects can be found not only in infinitive complement clauses, like in English, but they are also allowed in tensed constructions where the subject can be retrieved from the context. As pointed out in linguistic studies for Italian (Rizzi 1982, Burzio 1986) it seems that there is usually a parallelism between this null subject (pro-drop) property and the fact that verbal inflection assumes pronominal properties, in the sense that it is specified like a clitic (clitic is a verbal affix with pro-nominal properties specified with respect to such grammatical features as person and number) and as a clitic can be interpreted. In our annotation schema a null subject is represented by a trace that encodes the agreement features of the verb, such as number, gender and person; notice that null subjects in un-tensed clauses require a reference index (in the empty constituent, as noted before). In the example represented in Fig. 6 two traces have been inserted. The first trace ([1] t / NULL - 21.1) represents a null subject of a tensed clause (*siamo convinti di offrire...*) and three information are showed in the box associated with this trace: gender masculine (M), number plural (PLU) and person 3 according to the inflectional features of the verb (*siamo convinti*). The second trace (t / NULL - 23.1 [1]) represents a null subject of an infinitive clause depending from the tensed clause. This second trace is co-indexed with the first null subject by the reference index [1]. Similar solution has been adopted in (Moreno et al. 1999) for a Spanish treebank.

## 4.2 Coordination

Dependency formalisms exhibit obvious difficulties with constructions that, differently from most structures, not involve a head and some dependents, such as coordination. Privileging one of the two conjuncts as a head of the whole coordination our approach presents an asymmetric

representation of coordinative structures. The resulting structure is shown in the example represented in Fig. 7: the head of coordinative dependency relation is the first conjunct.

In this example two coordinative constructions occur; in the first coordination (*In Serbia e Bulgaria*) all conjuncts are syntactically complete. This is not the case of the second coordination, determined by the comma, where conjuncts are not strings of words which could occur with the same meaning without the coordinate structure. In this second coordination the first conjunct is syntactically complete (*...In Serbia e Bulgaria protestano contro gli ex comunisti...*) but in the second conjunct (*..., a Tirana contro un Governo di centro...*) the verb is gapped. This is a case of a relatively common phenomenon occurring in coordinative structures, called *gapping*.

In the Penn treebank the phenomenon of gapping is represented through a simple notational mechanism that allows the predicate-argument structure of gapped clauses to be recovered in most cases. In this formalism the complete clause is used as a template for the gapped one, together with a notation for mapping the gapped clause onto that template.

In our approach gapped elements are represented using null elements co-indexed with words in the complete clause. For example in Fig. 7, a trace (t / VERB-MAIN - 10.3 [!]) co-indexed with the main verb of the first conjunct ([!1] *protestano* / VERB-MAIN - 5), is inserted to represent the verb gapped in the second conjunct.

## 5. Conclusions

In this paper we have presented the Turin University Treebank project describing in particular the annotation schema and pointing out the main features of our approach. The annotation schema is a dependency-based format augmented with null elements and enriched trough

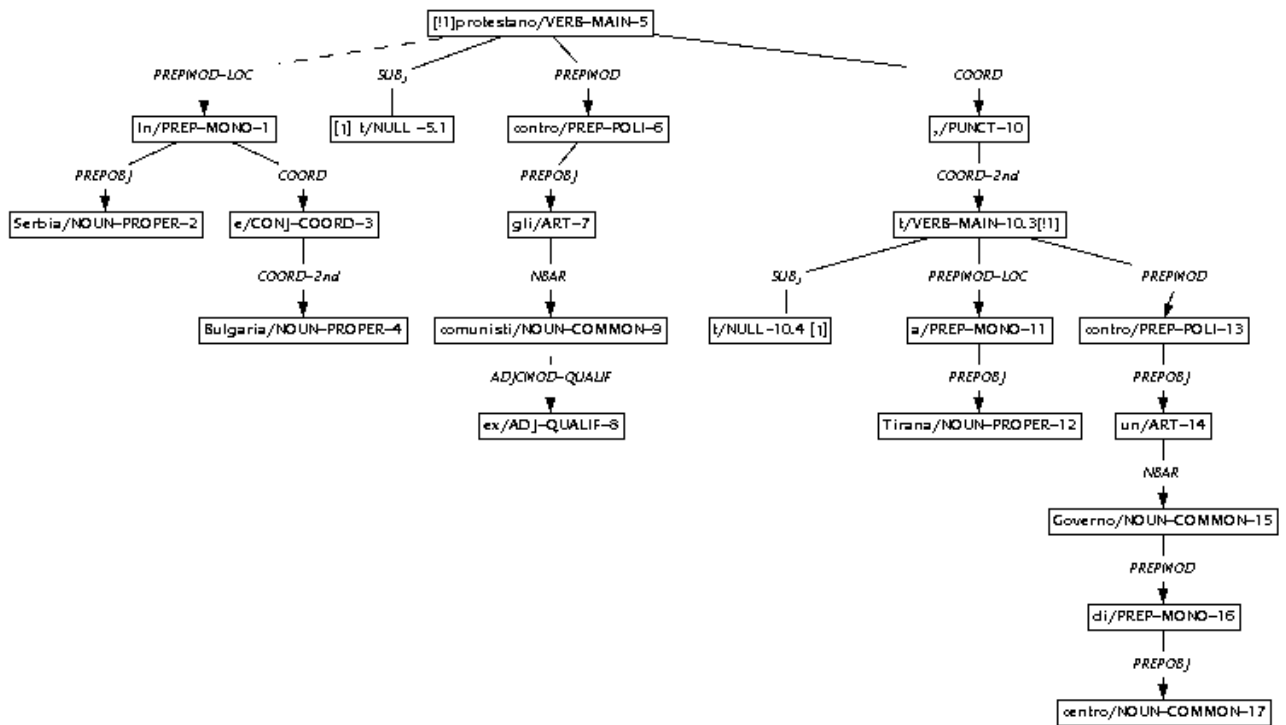


Figure 7 "In Serbia e Bulgaria protestano contro gli ex comunisti, a Tirana contro un Governo di centro" (In Serbia and Bulgary [they] protest against ex-communist, in Tirana against a Governement of the Centre)

a refinement of grammatical relations. The annotation of the treebank is a data-driven process performed by the help of an interactive parsing tool that build incrementally the syntactic representation of the sentence. We have also presented examples of representation of phenomena found in the corpus.

## 5. Acknowledgements

For the corpus to Cristina Cenci and Enrico Pozzi.

## 6. References

- ATALA (1999). Proceedings of Treebanks workshop – Journées ATALA sur les corpus annotés pour la syntaxe, 18-19 juin 1999, Paris, <http://talana.linguist.jussieu.fr/treebanks99/>.
- Bémová, A., Hajic, J., Hladká, B., Panenová, J. (1999). Syntactic tagging of the The Prague dependency Treebank. In Proceedings of Treebanks workshop - Journées ATALA sur les corpus annotés pour la syntaxe, 18-19 juin 1999, Paris, pp.21-29.
- Böhmová, A. & Hajicová, E. (1999). How much of the underlying syntactic structure can be tagged automatically?. In Proceedings of Treebanks workshop - Journées ATALA sur les corpus annotés pour la syntaxe, 18-19 juin 1999, Paris, pp.31-39.
- Boella, G. & Lesmo, L. (1998). Automatic Refinement of Linguistic Rules for Tagging. In Proceedings of 1st International Conference on Language Resources and Evaluation (LREC 98), Granada, pp.923-930.
- Brants, T., Skut, W., Uszkoreit, H. (1999). Syntactic annotation of a German newspaper corpus. In Proceedings of Treebanks workshop - Journées ATALA sur les corpus annotés pour la syntaxe, 18-19 juin 1999, Paris, pp.69-76.
- Brants, T. & Skut, W. (1998). Automation of Treebank Annotation. In Proceedings of the Conference on Methods in Language Processing (NeMLaP-3) January 14-17, 1998, Sidney, Australia, pp.49-58.
- Brants, T., Skut, W., Krenn, B. (1997). Tagging Grammatical Functions. In Proceedings of EMNLP-97, 1997, Providence, RI, USA, pp.64-74.
- Burzio, L. (1986). Italian Syntax. D. Reidel Publishing Company, Dordrecht/Boston/Lancaster/Tokyo.
- Church, K.W. & Mercer, R. L. (1994). Introduction to the Special Issue on Computational Linguistics Using Large Corpora. In Armstrong S. (ed.), Using Large Corpora, The MIT Press, Cambridge, Massachusetts, pp.1-24.
- Hajic, J. (1998). Building a Syntactically Annotated Corpus: the Praga Dependency Treebank In Issues of Valency and Meaning, Karolinum, Praha, pp.106-132.
- Hudson, R. (1984). Word grammar. Basil Blackwell, Oxford and New York.
- Hudson, R. (1990). English word grammar. Basil Blackwell, Oxford and Cambridge, MA.
- Lombardo, V., Lesmo, L. (in press). A formal theory of dependency syntax with non-lexical units. To appear in Traitement Automatique des Langues.
- Lombardo, V. & Sturt, P. (in press). Incrementality and lexicalism: a treebank study. To appear in Stevenson, S. & Merlo, P., (eds.) Lexical Representations in Sentence Processing, John Benjamins: Computational Psycholinguistics Series.

- Lombardo, V. & Lesmo, L. (2000). Automatic refinement of grammatical relations. In this volume.
- Lombardo, V., Bosco, C., Vassallo, D., Lesmo, L. (submitted). Treebank annotation and psycholinguistic modeling.
- Marcus, M.P., Santorini, B., Marcinkiewicz, M.A., (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19, pp.313-330.
- Marcus, M.P., Kim, G., Marcinkiewicz, M.A. (1994). The Penn Treebank: Annotating Predicate Argument Structure. In *Proceedings of The Human Language Technology Workshop*, San Francisco, Morgan-Kaufmann.
- Moreno, A. & Lopez, S. (1999). Developing a Spanish tree bank. In *Proceedings of Treebanks workshop - Journées ATALA sur les corpus annotés pour la syntaxe*, 18-19 juin 1999, Paris, pp.51-58.
- Rizzi, L. (1982). *Issues in Italian Syntax*. Foris publications, Dordrecht-Holland/Cinnaminson-USA
- Skut, W., Krenn, B., Brants, T., Uszkoreit, H., (1997). An Annotation Scheme for Free Word Order Languages. In *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP)*, Washington, D.C..
- Skut, W., Brants, T., Krenn, B., Uszkoreit, H., (1998). A Linguistically Interpreted Corpus of German in Newspaper Texts. In *Proceedings of 1st International Conference on Language Resources and Evolution (LREC 98)*, Granada, pp.705-713.
- Stock, O. (1989). Parsing with Flexibility, Dynamic Strategies, and Idioms in Mind. *Computational Linguistics*, Vol.15, Num.1, March 1989, pp.1-17.