A protocol to unify annotative standards on Italian *treebanks*

Maria Parascandolo, Francesco Cutugno

LUSI-Lab - Dipartimento di Scienze Fisiche - Università di Napoli "Federico II"

Abstract

It is well known that the actually Italian treebanks scenario is configured in a fuzzy and inhomogeneous way, with differences bound to the theoretical background as to each specific annotative project goals. For the present study we choose three Italian *treebanks*: TreSSI *Treebank sintattico-semantica dell'Italiano*, TUT, *Turin University Treebank* and AN.ANA.S, *Annotazione e analisi sintattica*. This work is articulated in three steps: at first, an analysis on the annotation schemes used in the examined *treebanks* is performed, then a comparison between them is shown and, finally, we define a new hybrid format in which we wanted to appear only the informations we deemed relevant. Thus we obtained an essential *tagset*, having just five tags and conveying the majority of the informations by the use of attributes, as stated in the metalanguage XML. This simplification didn't cause informational power loss, because the analytical details moved from the constituency tags to the attributes. As already seen in TUT, the standard we formalized is structurally similar to Penn. This standard definition, by allowing the merging of the textual data contained in the main Italian *treebanks*, can offer considerable advantages on the linguistic-descriptive level for those who wish to look firstly at the data as a source of possible theories. From a computational point of view it represents, nonetheless, a solution to merge into a single dataset, queriable with the same rules, the material contained in the considered Italian *treebanks*, improving the power and the portability of Italian linguistic resources.

Keywords: Italian, syntax, treebanks, Treebank unification, conversion tools

1. Introduction

In this paper we will present a proposal to unify syntactic annotations developed for the main Italian *treebanks* existing at present time: *Treebank sintattico-semantica dell'Italiano* (TreSSI: Lenci and Montemagni, 1999), *Turin University Treebank* (TUT: Bosco et al., 2000) e *Annotazione e analisi sintattica* (AN.ANA.S.: Voghera and Cutugno, 2003).

The goal of this work consists in merging linguistic material contained in these three datasets in order to build a considerably more extended and portable *treebank* resource. Having such a powerful resource, containing a major part of the syntactically annotated Italian textual material, will definitely help improving automatic instruments for automatic language processing and will allow us to optimize our capability to infer linguistic models from these databases. This proposal directly derives by a data-driven approach and, at this initial stage, it is not explicitly connected to a specific theoretical paradigm. The definition of a preliminary *tagset* (Tab. 6) is to be considered only an initial attempt to break the circularity (tags and grammatical categories always derive form theories!) by choosing a sort of non-empty intersection of elements meanly belonging to different pre-existing labelling sets and aiming at an inductive mining leading to the emergence of a formal description framework.

The solution we propose here defines a new hybrid syntactic *tagset* and takes advantage of the XML metalanguage which, thanks to its flexibility, allows to avoid the obstacles represented by

the informative power loss moving, as we will show, the analytical details from the constituency tags to the attributes. This way we obtained a syntactic *tagset* that, while being essential, is informationally rich and thus allowing us to prune the tree structure.

2. Annotative patters comparative analysis on TreSSI, TUT and AN.ANA.S.

2.1. Italian treebanks: the main issue

Building *treebanks* for the Italian language is a task that, in the last few years, has been thoroughly addressed, thus making possible for this language to emerge from the marginality situation in which it was confined. However, having different *treebanks* proliferate, due to different specific research projects, causes the overview on available Italian syntactic annotations to be very complex and diversified. This complexity is not functional for the needs of the computational linguists nor for the needs of those who develop applications in Natural Language Processing.

The main issue we face today for a language like Italian lies in the lack of a *treebank* of significantly extended dimensions and in the lack of uniform annotative rules.

Regarding the size, none of the Italian *treebanks* reaches, at present time, a dimension comparable, for example, to the *Penn Treebank* for the English language (Marcus et al., 1993): just as a reference for magnitude comparison, Italian *treebanks* contain, at best, hundreds of thousands tokens against the 3 million tokens contained in Penn, considering just the corpus portion that is syntactically annotated. Besides, heterogeneity in annotational formalisms, bounded to the difference of theoretical frameworks as to each annotative (annotative) project's specific goal, compromises the computational tractability of the Italian language.

Our study started from here, from this shape-shifting variety of annotative formalisms, trying to check if, beyond the evident differences existing between annotative patterns in Italian *treebanks*, a somehow shared nucleus was hidden.

In other words, this research tries to investigate the chances to standardize annotative formalisms used in the most extended Italian *treebanks*. Today it is crucial to face the issue of standardization because it represents the one and only way to unify the rules used to parse textual data, thus obtaining a *treebank* that is more extended and uniformly annotated. *Corpus Linguistics* is a discipline that works with big numbers, on the chance for linguistic phenomena models to be inferred from regular repetitions observed inside a corpus, and it would be quite difficult to handle big datasets should they be not uniformly structured because of different theoretical and methodological frameworks. Moreover, creating an automatic annotative process for textual material is made impossible by the variety of annotative formalisms.

Even though efforts to reach standardization are being made throughout the international community, these problems still represent a lightly explored field for the Italian language.

Our goal was, by punctually comparing the three annotative patterns, to obtain the definition of a new format in which we wanted to appear only the data we deemed relevant. This comparison, at present time, regards only the constituency-based releases of considered *treebanks*.

In a preliminary step, we prepared a comparative grid capable of clearly showing the peculiarities of the resources we examined against the parameters we considered relevant. These classification parameters, listed here, represent, in our opinion, the main varying dimensions found in the *treebanks* considered in this study regarding syntactic and morphosyntactic informations encoding.

- One-layer vs multi-layer representation (one/multi);
- Constituency annotation vs dependency annotation (const/dip);
- Annotation methodology (manual/semi-aut);
- Empty categories presence (+/-);
- Categorical *tagging* vs functional *tagging* (cat/func);
- Minimal representation (+/-min);
- POS *tagging* presence (+/-).

Applying these parameters to the selected *treebanks,* we obtained the classification shown in Tab. 1, where the value + means the feature is present in the corresponding annotative schemes, while the value – means it is not. Whenever one of the two values is written between brackets, that means it can be inferred from other available informations.

Parameters	*TreSSI*	*TUT*	*AN.ANA.S.*
One	+	–	+
Multi	–	+	–
Const	+	+	+
Dip	+	–	(–)
Manual	–	–	+
Semi-aut	+	+	–
Empty categories +/-	–	+	–
Cat	+	+	+
Funz	+	–	–
+/- Min	+	+	–
+/-POS *tagging*	+	+	–

Table 1: Main features of the three Italian treebanks considered

From the data reported in Tab. 1, it is clear which encoding aspects have not been treated uniformly in the considered annotation schemes and which have been. As we can see, there is nothing in common between the three datasets but the presence of a constituency-based format. Despite of the variety of the annotative solutions adopted in the examined *treebanks*, our effort was to reach, by punctually comparing the three annotation patterns, the definition of a new format in which only the data we consider salient appeared. This operation has been performed aiming at both computational goals and theoretical-linguistic ones, since this discipline is more inclined in giving attention to statistic data and, more generally, to the role played by probabilistic models in linguistic description. As a matter of fact, the definition of a unique format, by allowing the merging of the textual data contained in the main Italian *treebanks*, can offer considerable advantages on the descriptive level. These advantages are not to be underestimated by who, like us, whishes to look firstly at the data as a source of possible theories. Our main goal was to reach an annotation standard, a hybrid format holding only the features we deemed salient for our aims from the three considered schemes.

Our comparison is articulated in two main sections: in the first one, we describe the work on syntactic *tagsets* for each *treebank* and we propose a new hybrid *tagset*; in the second one, we discuss annotative criteria to use with the new *tagset* and we show simple rules to shift from each tree structure to the base structure we chose.

2.2. Comparing tagsets

Regarding TUT, we show in Tab. 2 the syntactic features we want to be stored by italicizing them.

Tag	*Constituent type*
S	*Simple or complex sentence. Infinitive implicit dependent clause.*
NP	*Noun Phrase*
ADJP	*Adjective Phrase*
VP	*Verb Phrase*
PP	*Prepositional Phrase*
ADVP	Adverb Phrase
SBAR	*Explicit dependent clause or relative clause*
*	Null Elements

Table 2: TUT syntactic tagset

As can be seen, we selected the S tag to be used in the new format both for sentence and for clause, obviously accepting recursion.

Tab. 3 shows the TUT functional tagset, from which we selected only the PRD tag. On the contrary, we chose to omit the TUT small *tagset* of semantic roles to make the new format one-layered as TreSSI and AN.ANA.S.

SBJ	Surface subject
LGS	Logical subject in passive sentences
PRD	*Predicative complement*

Table 3: TUT functional tagset

Excluding the semantic roles *tagset* is the first step towards a standardization attempt.
Furthermore, we are definitely interested in keeping the POS *tagging* from TUT.
As far as TreSSI concerns, we selected tags (italicized) from the list shown in Tab. 4.

Tag	*Constituent type*
F	*Sentence*
SN	*noun phrase*
SA	Adjective phrase
SP	*prepositional phrase*
SPD	prepositional phrase DI “of”
SPDA	prepositional phrase DA “by, from”
SAVV	adverb phrase
SQ	Quantified phrase
IBAR	verbal nucleus with finite tense
SV2	Infinitival clause
SV3	participial clause
SV5	Gerundive clause
FAC	Sentential complement
FC	coordinate sentence
FS	*subordinate sentence*
FINT	+wh interrogative sentence
FP	puntuaction marked, parenthetical or appositional sentence
F2	*Relative clause*
F3	Fragment clause
CP	dislocated or fronted sentential adjuncts
CP_INT	interrogative clause with adjuncts at left pherifery
COORD/costituente	coordination with coordinating conjunction as head
COMPT	Transitive complement
COMPIN	intransitive complement
COMPC	*copulative/predicative complement*
DIRSP	direct speech

Table 4: TreSSI syntactic tagset

In the same way, (see Tab. 5), we selected from AN.ANA.S. data conveyed by attributes we want to keep. Note that many omitted data are not lost, they are retrievable by the tree structure.

Tag	*Constituent*	*Attributes with corresponding values*
S	*Sentence*	Split (START | MID | END) Uniclausal (T | F) Number of clauses
F	Clause	*Type (M | DEP | NOM)* Number of phrases *Link (S_CONJ | S_PREP | NULL | REL)* *Arg (T | F)*
NP	*Noun Phrase*	Lexeme *Mw (T | F)* N (T | F) Arg (T) Cl (T) *Sub (T | F)* *Obj (T | F)* Det (T | F) Mod (T | F) *Position (PRE | POST | NULL)* Infra (T) Fr (0 | 1)
VP	*Verb Phrase*	Lexeme *Mw (T | F)* *Cop_ Vb (T | F)* *N_of_Arg (0 | 1 | 2 | 3)* *Sat (T | F)* Per (0 | 1 | 2 | 3 | 4 | 5 | 6) Mod (T | F) *Sub (T | F | NULL)* *Sub_ Type (N | PRON | O)* *Position (PRE | POST)* Fr (0 | 1)
PP	*Prepositional Phrase*	Prep Lexeme *Mw (T | F)* N (T | F) Arg (T) Cl (T) Det (T | F) Mod (T | F) *Position (PRE | POST | NULL)* *Modified Phrase (NP | VP | PP | PREDP | NULL)* Infra (T) Fr (0 | 1)
PredP	*Predicative Phrase*	Lexeme *Mw (T | F)* P_ of_ Speech (N | ADJ | PRON | O) Arg (T) Cl (T) Det (T | F) Mod (T | F) Position (PRE | POST | NULL) Fr (0 | 1)
DM	Discourse marker	Word Phrase Clause
COORD	Coordination	-------
ISO	Isolated	Type (ADV | ADJ | PREP | CONJ | N | V | PRON | ART | INT | PH)
HES	Hesitation	-------
REP	Repetition	-------
RR	Retrait-and-repear sequences	-------

Table 5: AN.ANA.S syntactic tagset

3. Hybrid *tagset* proposal

Firstly, we wish to point up some matches among the features we selected in the three *tagsets*.

Data conveyed by the SBAR label in TUT is the same that is held, in TreSSI, by two labels: F2 for relative clauses and FS for subordinate clauses. In our hybrid format, we don't need FS, because the information about subordination is held by the value of the S Type attribute. We also don't need F2 because, to mark a relative subordinate clause, the Link attribute holding the value REL is sufficient along, obviously, with the Type attribute value set to DEP. In the new syntactic *tagset* no specific tag will appear to mark subordinate clauses because we prefer to use, more simply, the S label along with its attributes values.

S	*Sentence* Type (M | DEP | NOM) Link (S_CONJ | S_PREP | NULL | REL) Arg (T | F)
NP	*Noun Phrase* Mw (T | F) Sub (T | F) Obj (T | F) Position (PRE | POST | NULL) Prd (T | F)
VP	*Verb Phrase* Mw (T | F) Cop_ Vb (T | F) N_of_Arg (0 | 1 | 2 | 3) Sat (T | F) Sub (T | F | NULL) Sub_ Type (N | PRON | O) Position (PRE | POST | NULL)
PP	*Prepositional Phrase* Mw (T | F) Position (PRE | POST | NULL) Modified Phrase (NP | VP | PP | AdjP | NULL) Prd (T | F)
AdjP	*Adjective Phrase* Mw (T | F) Position (PRE | POST | NULL) Modified Phrase (NP | VP | PP | NULL) Prd (T | F)

Table 6: Hybrid syntactic tagset

The COMP constituent in TreSSI marks complements of copulative verbs. We could say that it subsumes, in a way, the PredP constituent in AN.ANA.S. which is used to convey just the nominal part of the predicate. In our *tagset* we chose to omit both the tags, COMP and PredP, because data held by them can be recovered in an easier way. In fact, by merging the complements borne by copulative verbs inside VP, we can infer the data regarding the verb's copulative nature from the Cop_Vb attribute value. Regarding the nominal part of the predicate, we added the boolean attribute Prd, which was present in TUT functional *tagset* as one of the attributes belonging to NP and PP. This way, Prd value shall mark the nominal part of

the predicate while marking, at the same time, an eventual predicative nature of complements borne by other verb classes.

Our work mainly consisted in moving the analytical details from constituency tags to attributes. This way, we obtained an essential *tagset*, using just five labels and holding the major part of the data in the attributes, as stated in the standard marking metalanguage XML. The hybrid *tagset* is shown in Tab. 6.

It is straightforwardly clear that our *tagset* provides, by using a Boolean attribute, an explicit way to mark a multiword expression, whether it has a nominal, verbal or prepositional basis.

In fact, the only valid way to computationally work on the multiword dimension is, firstly, to perform a census of these structures and then, while annotating, to manually mark their presence using the Mw attribute.

3.1. Comparing tree structures

One of the most difficult issues that must be confronted when comparing annotative schemes is represented by the annotative criteria, namely the rules used to link the constituency labels to texts. In other words, the base structure of the tree that will be used must be specified. For simplicity and immediate computational tractability reasons, we chose the TUT tree structure. This structure, naturally, being similar to the one used by the Penn trees, is bounded to a configurational view of the syntax and to a massive presence of empty classes. Furthermore, it uses the *Chomsky-adjuction* to mark different phenomena: for example, in the annotation of verbs coming with there auxiliary, these are placed in a duplicated VP node at a higher level. Also, in the annotation of NP modifiers, they are linked to the highest NP node. Regarding noun phrases pre-modifiers, on the other side, TUT uses, like Penn, another principle by directly including them at same level of the head. Concluding, the TUT trees are characterized by their relative "flatness", because they present a minimal projection level and a noun phrases representation flattening.

Generally, these Penn-like annotative solutions, clearly implying a fairly strong compliance with the X-bar theory, reveal themselves to be surprisingly interesting, from our point of view, because of their actual descriptive power. We chose this kind of representation, even though we don't entirely share the same theoretical background, because the goal we are trying to achieve is just to describe linguistic phenomena and not to explain them, by adopting a particular theoretical framework.

When defining our annotational schema we tried to limit, when possible, the arbitrariness and subjectivity to which, inevitably, even a descriptive scheme is subject to.

On the other side, the TUT tree structure is less restrictive than the one used for Penn. In Penn, following the X-bar syntactic rules, every node must be binary, in the sense it cannot generate more than two branches and there are no cases violating this rule. Differently, the structure implemented in TUT accepts trees with nodes generating three or more branches. We think that the difference between a strictly binary representation and another one that is not, is, after all, irrelevant from a descriptive point of view.

One of the most relevant changes we made on TUT annotative scheme is definitely the empty classes removal along with the theoretical implications related to them. To balance the informational loss, this removal operation implies, in cases where constituents exhibit a non-canonical order, to add to the XML file some attributes, which were, by the way, already present in AN.ANA.S.

These attributes aim to specify the various constituents relative position. For NP, the Position attribute (PRE | POST | NULL) marks the NP position relatively to the verb; for VP, the Position attribute (PRE | POST) marks the position of the verb relatively to the subject; for PP, the Position attribute (PRE | POST | NULL) shows the PP position relatively to the phrase it modifies while the Modified Phrase attribute (NP | VP | PP | PREDP | NULL) holds the phrase type PP modifies.

Shifting back to the reference tree structure, it will obviously not be as complex as the TreSSI one is, but it will be adequately reduced until it converges on TUT.

Now we summarize the pruning steps needed to transform the base TreSSI tree shown in Fig. 1a into the one shown in Fig. 1b:

Prune the node FC/FS/FINT to its maximum projection. Data conveyed here can be moved, without informational power loss, under the S node of the tree in Fig. 1b;

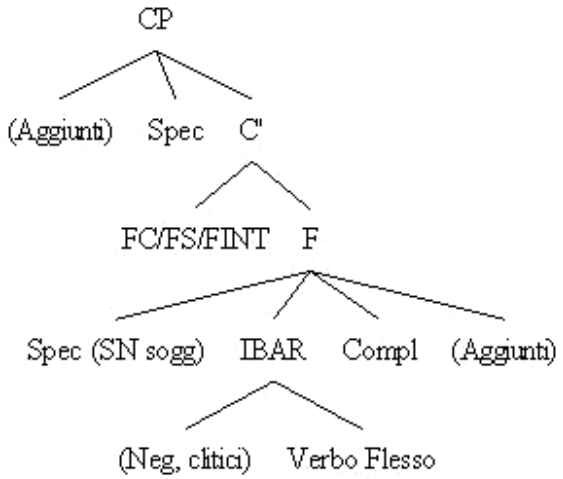

Figure 1 (a)

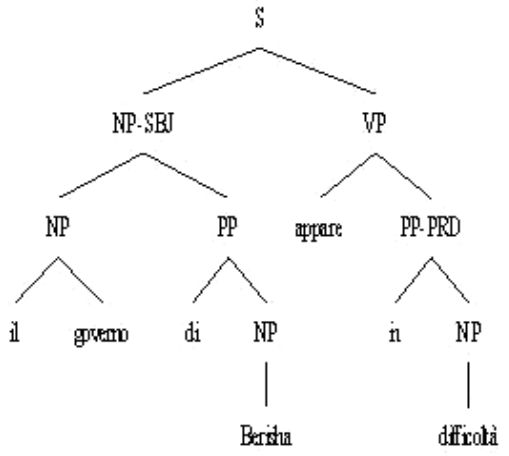

Figure 1 (b)

The F node of the tree in Fig. 1a overlaps S, thus becoming the unique node for clause and sentence;

The IBAR node overlaps the VP node;

The generalized Compl node can be pruned and embedded in VP.

3.2. Examples of sentence transformation

Original sentence:

La quota di azioni riservata al mercato statunitense è ancora top secret
The quote of shares reserved to market American is still top secret
e dipenderà ovviamente dagli umori del mercato alla vigilia dell 'Opv.
and (will)depend obviously from-the moods of-the market at-the eve of-the opv

Sentence annotated according to TreSSI guidelines

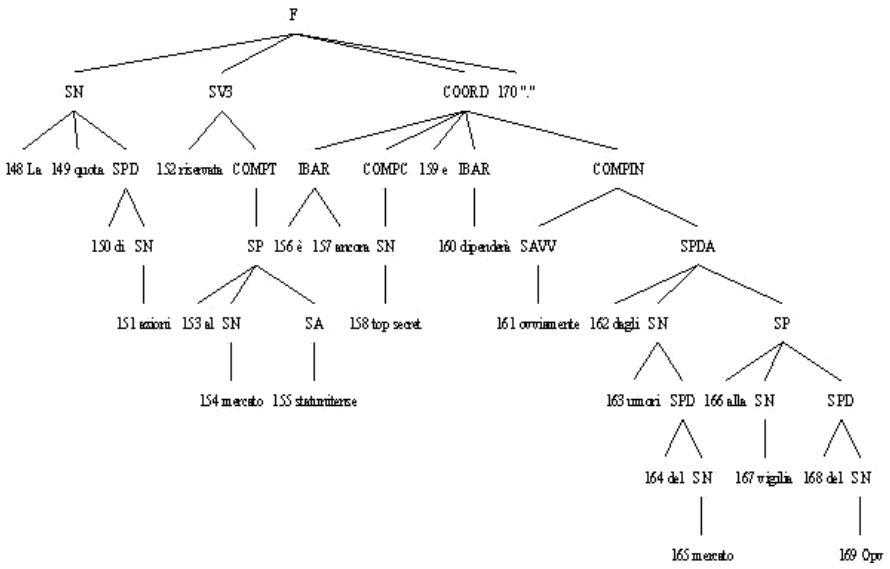

Figure 2: TreSSI annotation

Same sentence annotated according to TUT model:

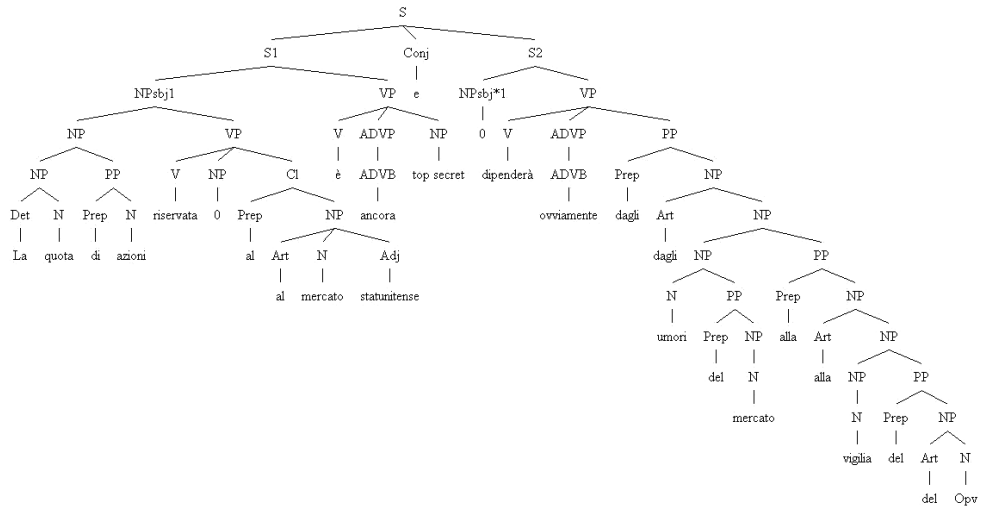

Figure 3: TUT annotation

Same sentence annotated according to AN.ANA.S model:

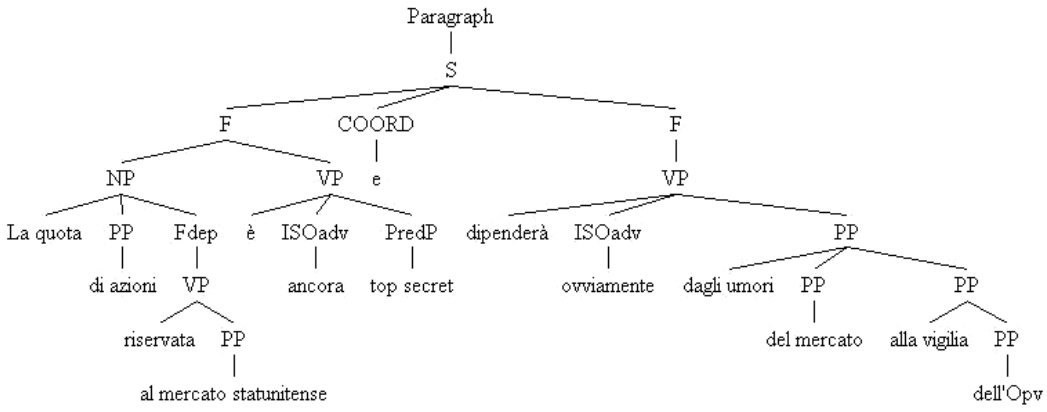

Figure 4: AN.ANA.S annotation

Same sentence annotated according to our hybrid model:

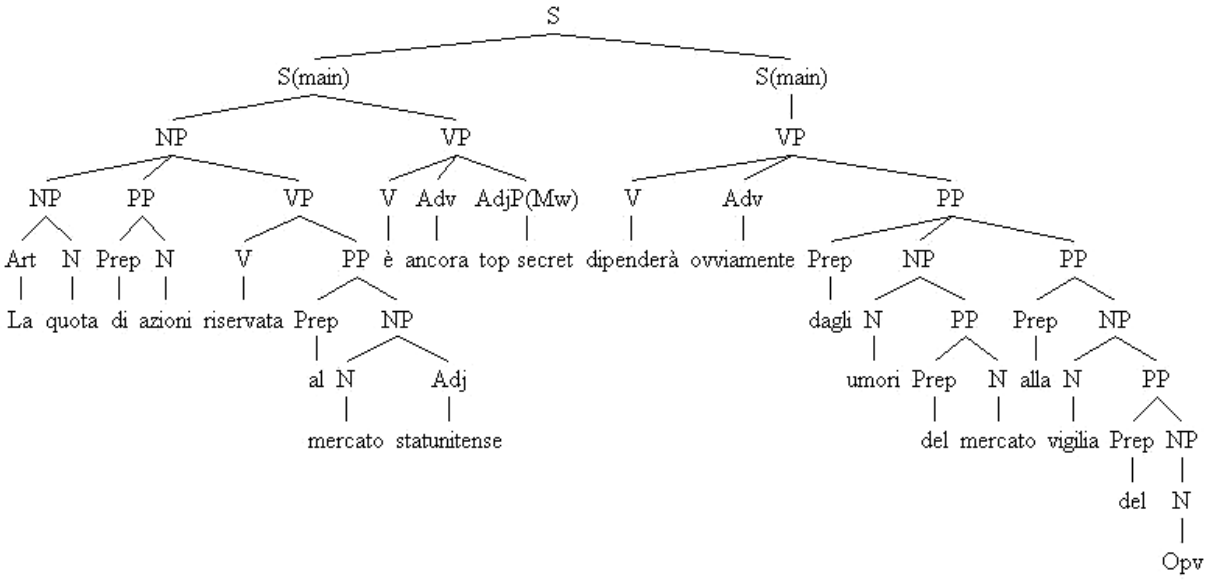

Figure 5: Hybrid annotation

The tree structure based on our hybrid model allows us to observe how we treat the adverbs. We do not presently provide constituency tag for this category. We presently choose to tag adverbs only with a pos label and to encapsulate them into the phrase they modify. In the presented example, the adverb *ancora* modifies the VP 'è' and the adverb *ovviamente* modifies the VP *dipenderà*.

In conclusion, our proposal differs in many aspects from the Penn-like class of models as it proposes the elimination of empty categories and the renounce to the binary edge structure.

Acknowledgements

We are grateful to Antonio Origlia for essential help and suggestions in translation.

References

Abeillé A. (2003). *Treebanks. Building and Using Parsed Corpora*. Dordrecht: Kluwer Academic Publishers.

Bosco C. (2003). *A grammatical relation system for Treebank annotation*. Ph.D. Thesis, Università degli Studi di Torino.

Bosco C. and Lombardo V. (2006). *Comparing linguistic information in treebank annotations*. In *Proceedings of LREC '06*, Genova.

Cutugno F. and Voghera M. (2004). *Analisi sintattica e annotazione XML a contatto*. In Albano Leoni, F., Cutugno, F., Pettorino, M. and Savy R., editors, *Il parlato italiano. Atti del convegno nazionale*, Napoli: D'Auria Editore.

Marcus M., Santorini B. and Marcinkiewicz M.A. (1993). Building a large annotated corpus of English: the *Penn Treebank*. *Computational Linguistics*, Vol. 19, 2: 313-330.

Montemagni S. (2001) La Treebank Sintattico-Semantica dell'Italiano di SI-TAL: architettura, specifiche, risultati. In *Atti del Workshop su "La Treebank sintattico-semantica dell'italiano di SI-TAL"*, *7° Congresso della Associazione Italiana per l'Intelligenza Artificiale (AI*IA 2001)*, Bari.