Treebanking in VIT: from Phrase Structure to Dependency Representation

Rodolfo Delmonte
University Ca' Foscari
Dept. Language Sciences - Laboratory Computational Linguistics

Abstract: In this chapter we will be dealing with treebanks, existing treebanks and their application fields. We will then describe VIT (Venice Italian Treebank), focussing on the syntactic-semantic features of the treebank that are partly dependent on the adopted tagset, partly on the reference linguistic theory, and, lastly - as in every treebank - on the chosen language: Italian. By discussing examples taken from treebanks available in other languages, we will show the theoretical and practical differences and motivations that lie behind our approach. In the end, we will discuss the quantitative analysis of the data of our treebank comparing them to other treebanks. In general, we will try to substantiate the claim that treebanking grammars or parsers is dramatically dependent on the chosen treebank; and eventually this process seems to be dependent both on substantial factors such as the adopted linguistic framework for structural description and, ultimately, the described language.

Keywords: Treebanks, syntactic representation, dependency structure, conversion algorithms, machine learning from treebanks, probabilistic parsing from treebanks

1. Introduction

In this chapter we will be dealing with treebanks, existing treebanks and their application fields. The questions that we ask ourselves are the following ones:
What's a Treebank? Which treebanks are there? Where are they - what languages? What dimensions and scope do they have? Are they on Written vs. Spoken Language? What types of linguistic representation do they use? What are their companion tools?
Treebanks have become valuable resources in natural language processing (NLP) in recent years. A treebank is a collection of syntactically annotated sentences in which the annotation has been manually checked so that the treebank can serve as a training corpus for natural language parsers, as a repository for linguistic research, or as an evaluation corpus for NLP systems. The course will serve as an introduction to the processes involved in creating and exploiting treebanks. We will give an overview of the annotation formats in different treebanks (e.g. the English Penn Treebank, the German TIGER treebank, the Venice Italian Treebank, etc.). We will

demonstrate important tools for the creation of treebanks (tree editors), for consistency checking in treebanks and for treebank searches. And we will look into the many usages of treebanks ranging from machine learning to system evaluation.
Creating a treebank from scratch is a hard task for a less studied language which in general lacks digitalized resources such as corpora onto which tagging has been carried out and checked manually. As will be argued in the sections below, this cannot be accomplished using freely available tools because they would require a tagged corpus. The suggestion is that of using Finite State Automaton to produce the rule set needed incrementally. One typical such tool for tagging is TBT – Transformation Based PoS Tagging - by Eric Brill(1995) or its correspondent Prolog version TnT by T.Brants (2000).
Uses for a treebank range from Parser Evaluation and Training; Parallel Treebanks for Machine Translation; Theoretical Linguistics validation and Grammar construction/induction.

2. Determining Factors in Treebank Construction

The following is a list of factors which we assume are of fundamental importance in deciding how the treebank and the underlying corpus should be organized. These factors are at the same time conditions of wellformedness of treebank and may constitute an obstacle against the usability of the same treebank for machine learning purposes. According to us, a treebank should be endowed with:

- Representativeness in terms of text genres
- Representativeness in terms of linguistic theory adherence
- Coherence in allowing Syntactic-Semantic Mapping
- Eventually highlight the distinctive linguistic features of the chosen language

Each factor can impact negatively on the linguistic texture of the treebank, and may thus undermine its utility in terms of general linguistic reference point for studies of the chosen language. In more detail, we assume that the factors above would be substantiated as follows:

- Corpus (Balanced) and representative of 6/7 different text genres vs. Unbalanced/Mono genre
- Strictly adherent to linguistic principles vs. loosely/non adherent (e.g. more hierarchical vs. less hierarchical)
- Constituency/Dependency/Functional structures are semantically coherent vs. incoherent
- Language chosen is highly canonical and regular vs. almost free word

order language

The final item is clearly inherent in the language chosen and not to be attributed to responsibilities of the annotators. However, as will be shown and discussed at length below, it may turn out to be the main factor in determing the feasibility of the treebank for grammar induction and probabilistic parsing.

2.1. Existing Treebanks and their main features

The main treebanks and related tools available nowadays are listed here below. They have been subdivided into 5 categories:
1. Feature Structure or Dependency Representation
2. Phrase Structure Representation
3. Spoken Transcribed and Discourse Treebanks
4. Tools
5. Other resources based on treebanks
6. Generic website for corpora

The full list is reported at the end of the chapter in an appendix. Next section will present in detail work carried out on the Italian treebank, which deals basically with syntactic representations. We will now briefly comment on the underlying problems of annotation and will focus in this section on discourse and semantic representation.

2.1.1. Annotating Discoourse and Semantic Structure

Treebank annotation is usually produced semi-automatically, but in the case of discourse and semantic representation it is only manual. Manual annotation is inherently an error-prone process so there is a need for very careful postprocessing and validation.
We can assume that beside syntactic trees, there are also two other types of similar hierarchical representation: semantic and discourse trees.
What do these tree represent? Depending on the theory behind it, we can think of the following:
- discourse structure is used to represent information about dependencies between units at the level of sentence/clause
- it is established on the basis of rhetorical relations, textual discourse dependence semantically founded, and eventually on the basis of communicative functions

Linguistic items relevant for the markup of discourse structure are all related to the notion of “coherence” and “cohesion”. They are:
- anaphoric relations
- typology of referring expressions
- discourse markers

As to theories supporting discourse and semantic representation we may assume the following are relevant ones:

- Intention driven (Grosz & Sidner, 1986)
  - Motivation for DS found in the intention behind the utterances
  - Discourse segments related by Dominance and Precedence
  - Tree structure constrains accessibility of referents

- Text Based (Mann & Thompson, 1988)
  - Motivation for DS found in the text
  - Discourse segments related on the basis of surface cues such as discourse markers
  - Relations between discourse segments labeled (eg. Elaboration, cause, contrast, etc.) from a finite - but potentially unlimited - set of DRs

- Discourse Information (Carletta et al. 1998)
  - Dialogue Tagging, intention based
  - Motivation for DS found in communicative functions
  - Segment labeled on basis of communicative intention
  - Restricted to three levels: Moves, speech acts; Games, goals; Transactions, topics;

These latter representations are not properly trees.

2.2. The theoretical framework

Schematically speaking, X-bar theory (we refer here to the standard variety presented in LFG theory) prefigures an organization of the type head and head-projections where each head is provided with a bar in hierarchical order: in this way the node on which a head depends is numbered starting from 0 and the subsequent dominant nodes have a bar, two bars and if necessary other bars (even though a two-bar projection is universally considered to be the maximum level). The hierarchical organization of the theory consists of the following abstract rewrite rules:

2.2.1. Theoretical scheme of X-bar rules

CP --> Spec, Cbar
Spec--> C0
C0 --> Complementizer
Cbar --> Adjuncts, XP
XP --> Spec, Xbar
Spec --> Subject NP
Xbar --> X, Complements/Adjuncts
X --> Verb, Adjective, Noun, Adverb

This rule schemata is however too weak to be of some use for practical purposes in a real corpus annotation task. So we operated a series of tuning and specialization of the X-bar schema while at the same time trying not to

betray its underlying foremost principle, which is the need that each constituent or higher projection should have only one and a single head.
Some decision taken was caused by the need to include under the same constituent label linguistic material belonging to the specifier which in our representation is only constituted by a positional variant: i.e. all constituents coming before the head are in the specifier of that constituent. The first choice we operated had to do with the internal organization of the specifier of NP that, in case of non-phrasal constituents, can consist of one or more linguistic elements belonging to different minor syntactic categories as reported below:

2.2.2. Atomic vs Structured Specifier

NP Spec--> Determiners, Quantifiers, Intensifiers
Verb Complex --> auxiliary verbs, modals, clitics, negatives, adverbials (also in a PP form), Verb

The choice to have a Spec structure was too difficult an option to pursue, so we decided to leave minor non-semantic constituents that stood before the head in an atomic form, unless it required a structure of its own, which could apply for quantifiers. Besides, semantic heads such as adjectives and adverbs always have their own constituent structure. As to the Verb Complex, it contained a number of atomic minor categories which we did not want to give a separate structure to if not required specifically. So, tensed verb takes a separate structure we have called IBAR - or IR_INFL ("unreal" verb) when the verb is either in future, conditional or subjunctive form- and that can consist of more elements added to the constituency level of the tensed verb.
Eventually we came up with the following less generic X-bar-like scheme:

2.2.3. X-bar rules for sentence level

CP --> SpecCP, Cbar
SpecCP -> Adjuncts, Fronted Complements, Focussed Arguments, Dislocated Constituents
Cbar --> C1, IP
Cbar --> C0, CP
C0 --> Complementizer
C1 --> Wh+ word

Here again it is apparent the need to specialize the rules: Cbar in case of wh+ words can never be followed by a CP, i.e. a subordinate clause starting with a subordinator. On the contrary, when a complementizer is instantiated CP may appear. The remaining rules are below:

IP --> SpecIP, Xbar, Complements, Adjuncts, Dislocated Constituents
SpecIP --> Subject

Complements --> COMPT/COMPIN/COMPC/COMPPAS
Xbar --> VerbalComplex
Spec --> Adverbials, Quantified Structures, Preposed Constituents

2.3. Syntactic Constituency Annotation

Eventually what we wanted to preserve was the semantic transparency of the constituency representation in order to facilitate the syntax-semantics mapping if needed. In particular we wanted the CLAUSE or IP to remain the semantically transparent syntactic nucleus corresponding to a Semantic Proposition with PAS. To that purposed we introduced a distinction between Tensed and Untensed Clauses, where the second need the unexpressed Subject to be bound to some Controller in the matrix clause. We were also obliged to introduce specialized constituency label by the specific features of the corpus we analysed: in particular, the texts are full of Fragments or Non-verbal sequences of constituents making a sentence.
Other specialized structures will be discussed further on, but now it is important to note that our representation does not employ a VP structure level: in fact, we preferred to analyse verbal group as directly positioned on the same level of S, where there will also be a NP-Subject, if syntactically expressed. We also decided to introduce a label for each of the three main lexical types specifying the syntactic category of the verbal governor to the complement structure which would thus be subcategorized according to different types of complements, among which we also introduced Voice or Diathesis to specialize the complements of a passive verb – COMPPAS, in order to allow an easy automatic conversion in case of the presence of an adjunct containing an agent in SPDA form. By doing this, VIT partially followed NEGRA, the German treebank, also in the sense of specializing major non-terminal constituents, as discussed in the sections below. While on the contrary PennTrebank (hence PT) differs for a less detailed and more skeletal choice, as specified in the PT guidelines. We show two examples below of how a structure in PT could be better represented using our rule schemata:

(1) In exchange offers that expired Friday, holders of each $1,000 of notes will receive $250 face amount of Series A 7.5% senior secured convertible notes due Jan. 15, 1955, and 200 common shares.

```
(
  (S (PP-LOC In
       (NP (NP exchange offers)
                 (SBAR (WHNP-1 that)
                                (S (NP-SBJ *T*-1)
                                   (VP expired
                                     (NP-TMP Friday))))))
    ,
    (NP-SBJ (NP holders)
```

```
        (PP of
          (NP (NP each $ 1,000 *U*)
            (PP of
              (NP notes)))))
    (VP will
      (VP receive
        (NP (NP (NP (ADJP $ 250 *U*) face amount)
                  (PP of
                    (NP (NP Series A
                                  (ADJP 7.5 %)  senior secured convertible notes)
                  (ADJP due
                    (NP-TMP (NP Jan. 15)
    ,
                                                                  (NP 1995))))))
                                  and
          (NP 200 common shares))))
    .)
 )
```

As can be easily noticed, the sentence S begins with an Adjunct PP – an adjunct NP would have been treated the same way – which is then followed by the NP subject always at the same level. In our representation, the adjunct would have been positioned higher, under CP,

```
(
  (CP (PP-LOC In
        (NP (NP exchange) offers
                  (CP (WHNP-1 that)
                                (S (IBAR expired)
                          (COMPIN (NP-TMP Friday))))))
    ,
     (S
        (NP-SBJ (NP holders
              (PP of
                (NP (QP each) $ 1,000 *U*
                 (PP of
                  (NP notes))))))
        (IBAR will receive)
        (COMPT (COORD
              (NP (NP (ADJP $ 250 *U*)
                    face amount
                     (PP of
                       (NP (NP Series A
                                          (ADJP 7.5 %)
                                          (ADJP senior secured convertible)
                           notes)
                          (ADJP due
                           (NP-TMP (NP Jan. 15)
                                                                  ,
                                                                  (NP 1995))))))
                                  and
                  (NP 200 common shares)))))
.)
```

Also notice that we add an abstract COORD node that in this case is headed by punctuation conjunction AND, and in other cases will be added by punctuation marks.
An interesting question is constituted by the role played by Auxiliaries in case they are separated from the main verb by the NP Subject, has happens in English and Italian with Aux-To-Comp structures, and in general in German with Verb Second phenomena which are very frequent. NEGRA treebank has solved the problem by inserting a special label at S and VP level as shown here:

```
(
 (S  (S-MO
     (VMFIN-HD Mögen)
     (NP-SB
              (NN-NK Puristen)
                 (NP-GR
                  (PIDAT-NK aller)
                  (NN-NK Musikbereiche) ))
     (ADV-MO auch)
     (VP-OC
              (NP-OA (ART-NK die)
                                  (NN-NK Nase) )
               (VVINF-HD rümpfen) ))
   ($, ,)
   (NP-SB (ART-NK die)
              (NN-NK Zukunft)
                 (NP-GR (ART-NK der)
                                  (NN-NK Musik) ))
   (VVFIN-HD liegt)
   (PP-MO (APPR-AC für)
              (PIDAT-NK viele)
              (ADJA-NK junge)
              (NN-NK Komponisten) )
   (PP-MO
              (APPRART-AC im)
              (NN-NK Crossover-Stil)
              ))
 ($. .) )
```

Having a more specialized inventory of constituents was done also in view of facilitating further conversion projects into dependency structure which will be illustrated below. It also allows for easy searches and better specification of the structure searched. In particular, having a specialized node for tensed clauses, which is different from the one assigned to untensed ones, allows for better treatment of such constituent, which, as will be shown below, allows for some of its peculiar properties to be easily detected. Moreover, by assuming that the tensed verb complex – IBAR/IR_INFL - is the sentence head is in line with a number of theoretical frameworks and allows a much easier treatment in the LPCFG (Lexicalized Probabilistic Context-Free Grammars) scheme, where the head of the VP is also the head of S. Differently from what happens with PT, in VIT it doesn't have to be

extracted from a substructure because it's already at S level: on the contrary, in PT the head could be the leaf of many different VP nodes depending on how many auxiliaries or modals precede the main lexical verb. In our case, for every further operation of transduction in dependency structure, the number of levels to keep under control is lower when the task of detecting Head-root and Head-dependent relations.

Adding a VP node that encompasses the Verbal complex and its complement was not a difficult task to carry out. We have then produced a script that enables the transformation of the entire VIT without a VP node into a version that conversely has it, but only in those cases where it is allowed by the grammar. In this way we successfully removed all those instances where the verbal group IBAR/IR_INFL is followed by linguistic material belonging to the S level, such as phrasal conjunctions, PP adjuncts or parenthetical structures. By doing this we were able to identify about 1000 clauses out of the total 16000 where the VP node hasn't been added.

The following section describes work carried out to produce an algorithm for the automatic conversion of VIT, which uses traditionally bracketed syntactic constituency structures, into a linear word-based head-dependent representation enriched with grammatical relations, morphological features and lemmata. We are also still trying to produce a machine learning parsing algorithm that performs better than the current accuracy results which range below 70%.

3. A Case Study: VIT – Venice Italian Treebank

The VIT Corpus consists of 60.000 words of transcribed spoken text and of 270.000 words of written text. In this chapter we will restrict our description to the characteristics of written texts of our Treebank, even though we will use the quantitative data of the spoken texts for comparisons with the written one.

The first version of the Treebank was created in the years 1985-88 with the contribution of Roberto Dolci, Giuliana Giusti, Anna Cardinaletti, Laura Brugè, Paola Merlo who also cooperated in the creation of the first Italian subcategorized frequency lexicon where the first 4.000 words in the frequency list of LIF were chosen. These procedures had been promoted by means of a research program financed by DIGITAL Equipment that was interested in building an Italian version of its voice synthetizer DECTalk, i.e. a system of vocal automatic synthesis from a written text in Italian based on the one realized for American English. To this end, it was necessary to recreate the same linguistic tools of the original version: that is a robust syntactic parser for unrestricted text, a morphological analyser and a lexicon that could work with unrestricted Italian texts without vocabulary limitations. The treebank created at that time was only in paper form, because of the lack of other samples available worldwide – the one created by the University of

Pennsylvania was a work-in-progress – and also for the lack of adequate software to produce annotation interactively and consistently.

The paper documents – that are still kept in the Laboratory of Computational Linguistics where they were produced – were used for the creation of a probabilistic context-free grammar of Italian, i.e. a list of all the rewriting rules produced by manual annotation and for every different rule the frequency value of the rule itself in the corpus. The chosen corpus consisted of 40.000 words taken from newspaper or magazine articles pertaining to politics, economics, current events and burocratic language: the texts were digitized and available on mainframe computers, but not annotated as for PoS. This phase of the work is documentated in a paper (Delmonte R. & R.Dolci, 1989). Work for the creation of the treebank was then discontinually carried on reusing the above-mentioned texts and gradually expanding the sample. This went on until the approval of the national project SI-TAL in 1998 which was also the right prompt to achieve a normalization of the overall syntactic annotation. The actual treebank uses those texts and others elaborated for the national project SI-TAL and the projects AVIP/API/IPAR as well as texts annotated on a number of internal project - as for instance one with IRST concerned with literary Italian texts.

The creation of a treebank is the last step in a long and elaborated process during which the original text undergoes a total transformation. The texts have been digitized and, if necessary, corrected – in case of orthographic or other sorts of errors, which have been removed in order to avoid unwanted and malformed syntactic structures. Subsequently, by employing the suite of automatic annotation programs by Delmonte et al.(2004), we proceeded to the tokenization of the texts, providing each word with a line or record and one or more indexes – in case the word was an amalgam or a cliticized verb. In this stage, we verified that those words consisting of a combination of letters and digits, letters and graphical signs, dates, formulas and other orthographic combinations that are not simple sequences of characters had been transformed appropriately and that no word of the original text had gone missing during the process.

From the resulting tokenized text we move on to the creation of Multiwords – more details in the following sections. This operation is accomplished using a specialized lexicon which has been created on purpose and in which one could add other forms or idiomatic expressions that have to be analyzed syntactically as one word because they constitute a single meaning and no semantic decomposition would be allowed. Inflected versions of each multiword had to be listed if needed.

In this stage we created a lexicon specialized to particular domains. This has been done in the case of the spontaneous dialogue texts of the national projects AVIP/API/IPAR (see Delmonte et al., 2004) where coding of semi-words, non-words and other forms of disfluencies has taken place; where possible the specific lexicon also contains reference to the lemma of the wordform.

Tagging is performed by assigning to each token previously found the tags or PoS labels on the basis of a wordform dictionary and of a morphological analyser that can proceed to do "guessing" in case the corresponding root cannot be found in the root dictionary – but see also Chapter 9. This operation is done by decomposing a word in affixes, inflections and derivational ones, in order to identify an existing root; in lack of such information, a word will be categorized with the temporary tag "npro" (proper noun) if uppercase or "fw" (foreign word) if lowercase. In this stage amalgamated words (e.g. DEL = Di/prep, lo/art_mas_sing), are split and two separate words are created; in addition to that, an image of the text in the form of sentences is created and these sentences will then be used for syntactic analysis which assumes the sentence as the ideal span of text. As already stated above, all steps of morphological analysis and lemmatization together with the creation of specific lexica and phase in which we built and analysed the multiwords have required one or more cycles of manual revision.

Tagging was completed by the semi-automatic phase of disambiguation, i.e. choice of single tag associated to every word according to context. The texts we analyzed showed on average 50% ambiguity level: this means that every word was associated to two tags on average. To solve the problem of word disambiguation we used hybrid algorithms that are in part statistical and in part syntactical and converge in a program that has an interface for the annotation which allows the annotator to take quick decisions as to which tag to assign in the actual context even when the correct tag differs from the ones displayed by the automatic analysis. In this way, the annotator has also taken care of those cases in which the system did not have enough lexical or morphological information to process the current word.

Eventually, parsing takes place. The automatic analysis of the parser is submitted to a manual check and in the end to the collation from a supervisor who is responsible of the eventual unification of the structural "variants" suggested by different annotators for the same structural type (there were only two). This operation was critical and has required in some cases a total revision of some parts of the treebank itself, as has been the case with comparative and quantified structures in the project SI-TAL (see Delmonte, 2004).

3.1. From Constituent Structure to Head-Dependent Functional Representation

This section describes work carried out to produce an algorithm for the automatic conversion of VIT, which uses traditionally bracketed syntactic constituency structures, into a linear word-based head-dependent representation enriched with grammatical relations, morphological features and lemmata.

Dependency syntactic representation consists of lexical items – the actual words - linked by binary asymmetric relations called dependencies. As Lucien Tesnière formulated it (1959):

> La phrase est un ensemble organisé dont les èlèments constituants sont les mots. Tout mot qui fait partie d'une phrase cesse par lui-meme d'etre isolé comme dans le dictionnaire. Entre lui et ses voisins, l'esprit aperçoit des connexions, dont l'ensemble forme la charpent de la phrase. Les connexions structurals établissent entre les mots des rapports de dépendance. Chaque connexion unit en principe un terme supérieur à un terme inférieur. Le terme supérieur reçoit le nom de régissant. Le term inférieur reçoit le nom de subordonné. Ansi dans la phrase "Alfred parle" ... parle est le régissant et Alfred le subordonné.

If we try to compare types of information represented by the two theories we end with the following result:

- Phrase structure explicitely represent Phrases (nonterminal nodes); Structural categories (nonterminal labels). Possibly some functional categories (grammatical functions)
- Dependency structures explicitely represent Head-dependent relations (direct arcs); Functional categories (arc labels). Possibly some structural categories (POS).

Some theoretical framework besides the founder of the theory are the following ones: Word Grammar (WG) Hudson 1984; 1990. Functional Generative Description (FGD) Sgall et al. 1986. Dependency Unification Grammar (DUG) Hellwig 1986;2003. Meaning Text Theory (MTT) Mel'cuk 1988. (Weighted Constraint Dependency Grammar (WCDG) Maruyama 1990, Haarper & Hazelman 1995, Menzel & Schroeder 1998, Schroeder 2002. Functional Dependency Grammar FDG Tapanainen & Jaervinen 1997, Jaervinen & Tapanainen 1998. Topological/Extensible Dependency Grammar (T/XDG) Duchier & Debus 2001, Debusmann et al. 2004.

In short, we can define dependency syntax to have to the following distinctive properties:

- It has direct encoding of predicate argument structure
- dependency structure is independent of word order
- for that reason it is suitable for free word order languages (Latin, Walpiri, etc.)
- however, it has limited expressivity
    - every projective dependency grammar has a strongly equivalent context-free grammar but not vice-versa
    - impossible to distinguish between phrase modification and head modification in unlabled dependency structure

To obviate to some of the deficiencies of the dependency model, we designed our conversion algorithm so that all the needed linguistic information was supplied and present in the final representation as discussed in the section below.

3.2. The Conversion Algorithm

Input to the Algorithm for Head-Dependency Structures (hence AHDS) are the original sentence-based bracketed syntactic constituency structures, which are transformed into Head-Dependent column-based Functional representation by a pipeline of algorithms or rather scripts. These scripts produce a certain number of intermediate files containing the Tokenization, the Head Table, and the Clause Level Head-Dependency Table (hence CLHDT). The final output should be a file that contains the following items of linguistic information – for the word competitività/competitivity - in a column-based format:

id_num.	word	POS	role	id_head	const.	lemma
[semantic/morphological features]						
5	competitività	N(noun)	POBJ	4	SN	competitività
[sems=invar, mfeats=f]						

In the Tokenization file VIT is represented as a vertical list of words in the form of word-tag pairs. In addition, all multiword expressions have been relabeled into a set of "n" words preceding the head tagged as "MW". The Head Table defines what category can be head to a given constituent and also the possible dependents in the same structure. The Head Table differentiates dependents from heads and has been used together with the Tokenization file to produce the CLHDT file. The current Tokenization includes information as to the constituent label the category belongs to. It also differentiates between simple POS labels and rich labels with extended linguistic (syntactic, semantic, morphological) information.

The fully converted file also includes Grammatical Relation labels. In order to produce this output, we had to relabel NP SUBJects, OBJects and OBLiques which are placed in a non canonical position. A similar question is related to the more general need to tell arguments and adjuncts apart for ditransitive and intransitive constructions. In Italian, prepositional phrases can occur quite freely before or after another argument/adjunct of the same predicate. Our strategy was at first that of marking as OBLique the first PP under COMPIN, and of course PPby under COMPPAS (more on this in a section below). But there is no possibility to mark ditransitive PP complement without subcategorization information, nor for that matter would PPs marked OBLiques be fully compliant without lexical information.

The solution to this problem was on the one side the use of our general semantically labeled Italian lexicon which contains 17000 verb entries together with a lexicon lookup algorithm, where each verb has been tagged with a specific subcategorization label and a further entry for prepositions subcategorized for. The use of this lexicon has allowed the automatic labelling of PP arguments in canonical position and reduced the task of

distinguishing arguments from adjunct to the manual labeling of arguments in non canonical position.

On the other side, seen that nominal heads have been tagged with semantic labels – see the tagset in the appendix -, we proceeded at first by labeling possible adjuncts related to space and time. In case of verb of movement, where the subcategorization frames required it, and the preposition heading the PP allowed it, we marked the PP as argument. We also relabeled as arguments all those PPs which were listed in the subcategorization frames of Ditransitives, again where the preposition allowed it.

We organized our work into a pipeline of intermediate steps which were incrementally turned to the full conversion task. In this way we also managed to check for consistency at different levels of representation.

3.3. Tagging and Multiwords

Checking consistency at the level of categories or PoS, was the work done with the first step, the tokenization of the VIT. At this level, we had to recover full consistency with multiwords as they had been encoded in the current version of VIT. We are aware of the fact that the lack of such an important annotation has caused serious problems in the PennTreebank where the same problem has been solved by assigning two different tags to the same word: e.g. the word "New" is tagged NNP and not JJ if it is followed by another NNP - "York" for example –, to convey the fact that "New" has to be interpreted as part of the proper name "New York". However this has no justification from a semantic point of view seen that "New York" as a geographical proper name needs to use both words in order to access its referent not just one. Perhaps the original meaning of the word "New" in "New York" was that of adjective (hence JJ), seen that "York" in the new continent was "new" in relation to the British corresponding city name. But of course, all those words that encode their meaning in more than one wordform, will not be captured as such in the Penn treebank. For sure, the use of NNP for the non semantically independent portion of proper names will only contribute ambiguity to the same wordform that in other context will be tagged with their "natural" literal meaning. The conversion process starts with The script takes the parenthesized VIT as input file and creates a treebank version with indices without words and then the complete head table where every constituent is associated to its head with word id. To get that we differentiate nonterminal symbols from terminal one and assign an incremental index number to the latter.

As shown in Table 2. we eventually produced a verticalized version which contains PoS labels and their fully specified meaning, followed by constituent label in which the word was contained. In addition, PoS labels have been commented and whenever possible, morphological features have been added.

3.3.1. Head-constituent relations

As a second step in our work we produced the table of Heads/constituents relations according to the rules formulated below. This step obliged us to look into every relation carefully so that no category was left without a function: it could either be a dependent or a head. No dangling categories would be allowed. We discovered that in the case of comparative constructions there was the need to separate the head of the phrase from the second term of comparison which did not have any specific constituent label. Working at constituent level we have been then obliged to introduce a new constituent label SC for comparative nominal structures, a label which is also used for Quantified related constructions. Rules are specified in the table below. The head extraction process was carried out following a list of head rules – some of which are presented here below - according to Collins' model for English. In particular, *Direction* specifies whether search starts from the right or from the left end of the child list dominated by the node in the *Non-terminal* column. *Priority* gives a priority ranking, with priority decreasing when moving down the list:

Table 1. Head-Constituent relations

Constituent Non-terminal	Direction	Priority list
AUXTOC	Right	ause, auag,aueir,ausai,vsup
SN	Right	n,npro,nt,nh,nf,np,nc,sect,fw,relq,relin,relob,rel,pron, per_cent,int,abbr,num,deit,date,poss,agn,doll,sv2,f2,sa, coord
SAVV		part,partd,avvl,avv,int,rel,coord,fw,neg,f2
SA	Right	ag,agn,abbr,dim,poss,neg,num,coord,ppre,ppas,fw,star,f2
IBAR	Right	vin,viin,vit,vgt,vgin,vgc,vppt,vppin,vppc,vcir,vcl,vcg,vc, vgprog,vgsf,virin,vt,virt,vprc,vprin,vprogir,vprog,vprt,vsf ,vsupir, vsup,vci,coord

3.4. Clause Level Head-Dependency Table (hence CLHDT)

The third step in our work has been the creation of the CLHDT which contains a column where word numbers indicate the dependency or head relation, with the root of each clause bearing a distinctive dash, to indicate its role, as shown in Table 3. Rules for head-dependent relations are formulated below.

3.4.1. Rules for Head-Dependent Relations

At first we formulated a set of general rules as follows:
Heads with no constituent – or dangling heads - are unallowed.
Constituents with no heads are unallowed.

Coordinate structures are assigned an abstract head: they have conjunctions, punctuation or nil as their heads. Conjunctions are a thorny question to deal with: in dependency grammars they are not treated as heads. However, we take this case to represent a simple case of functional head government, very much in vein with a complementizer heading its complement clause. Punctuation plays an important role in parsing and in general it constitutes a prosodically related non-linguistic item. This is very clear in transcribed spoken corpora where all pauses had to be turned into appropriate punctuation, as we had to do in our work on Italian Spontaneous Speech Corpora (see Delmonte et al., 2007). This is why we assign a similar treatment to all "meaningful" punctuation marks. Punctuation marks like dash, quotations, parenthesis, angled brackets, which may introduce Parentheticals, Direct Speech, Reported Direct Speech are treated as functional heads. Other punctuation marks like commas introduced just to mark a pause and play no additional structural role are left interspersed in the text, as happens with PTB.

To better grasp the role of each constituent and its head in the conversion task, we divided up constituents according to their function and semantic import, into three main categories. As can be noticed, we specialized our non generic X-bar scheme into a set of constituent labels which were required to set apart functional types as well as structural and semantic types. For these reasons sentential constituent typologies differentiate between:

- Simple Declarative (F)
- Complex Declarative (CP)
- Subordinate Clause (FS)
- Coordinate Clause (FC)
- Complement Clause (FAC)
- Relative Clause (F2)
- Nonfinite tense Clause (SV2-SV3-SV5)
- Interrogative (FINT, CP_INT)
- Direct (Reported) Speech (DIRSP)
- Parenthetical, Appositive and Vocative (FP)
- Stylistically (literary and burocratic) marked utterances (TOPF)
- Fragments or non propositionally relatable utterances – lists, elliptic linguistic material, etc. (F3)

#ID=sent_00002		
F Sentence	**=**	**IBAR Verbal group with tensed verb**
COORD Coordinate structure for constituents	**=**	**CONG Conjunction**
SN Nominal phrase	**=**	**N Noun**
SN Nominal phrase	**=**	**N Noun**
SPD Prepositional phrase with preposition DI	**=**	**PD Preposition_di**
SN Nominal phrase	**=**	**N Noun**
SA Adjectival phrase	**=**	**AG Adjective**

IBAR Verbal group with tensed verb	**=**	**VC Verb_copulative**
COMPC Complements governed by Copulative Verbs	**=**	**SAVV Adverbial phrase**
SAVV Adverbial phrase		**= AVV Adverb**
SN Nominal phrase	**=**	**N Noun**
SPD Prepositional phrase with preposition DI Preposition_di_plus_article	**=**	**PARTD**

Tab. 2 Local Heads/Constituents Relations

3.5. Rules for Grammatical Relation Labels

The final step in the overall treebank full-fledged conversion is constituted by the assignment of Grammatical Relation labels/roles. In a language like English which imposes strict position for SUBJect NP and OBJect NP, the labeling is quite straightforward. The same applies for German which in addition has case marking to supplement for constituent scrambling, i.e. the possibility to scramble OBJect and Indirect OBJect in a specific syntactic area.

Differently from these two languages and other similar languages which constitute the majority of Western language typology, Italian is an almost "free word-order" language. In Italian, non canonical positions would indicate the presence of marked construction - which might be intonationally marked - as containing linguistic information which is "new", "emphasized" or otherwise non thematic. Italian also allows freely the omission of SUBJect pronouns whenever it is a discourse topic; it also has lexically empty non-semantic expletive SUBJects for impersonal constructions, weather verbs etc. This makes the automatic labeling of complements or arguments vs. adjuncts a difficult task to achieve, if tried directly from constituent labels without help from any external additional (lexical) information.

We thus started to relabel non-canonical SUBJect and OBJect NPs, but the idea was that of relabeling all non-canonical arguments. However, we realized that we could operate a distinction between SUBJect and complements in general, where the former can be regarded EXTernal arguments, receiving no specific information at syntactic level from the governing predicate to which they are related. On the contrary, arguments which are complements are strictly INTernal and are directly governed by the predicate, be it Verb, ADJective or Noun. Preposition constitute a case "per sé" in that they govern PPs which are exocentric constituents and are easily relatable to the NP head they govern. However, PPs need to be related to their governing predicate which may subcategorize for them or not according to Preposition type.

We thus produced rules for specific labeling and rules for default labeling. Default labeling is a generic less specific complement label which will undergo modification, if needed in the second phase. On the contrary, specific labeling will remain the same.

In more detail we carried out the following steps. First, we manually listed all s_dis (preposed subject under CP), s_foc (focalized object/subject in inverted position, no clitic), s_top (topicalized subject/object to the right, with clitic) and ldc (left dislocated complement, usually SA/SQ/SN/SP/SPD/SPDA).
Second, we compared all verbs to an external verb list with verb valence and assigned the OBL role to the prepositions heading an oblique constituent. Then, we assigned a semantic role to the head of every constituent according to the following rules – we only list some of them:

Constituent	**Dependency**	**Role**
CCONG/ CONGF/ CONJL CCOM/ CONG	**Always**	**CONG**
SN/SQ	**Governed by F**	**SUBJ**
	Root of a sentence without a verb	**SUBJ**
	Governed by COMPT	**OBJ**
	Governed by COMPIN	**ADJ**
	Governed by COMPC	**NCOMP**
	Governed by F2	**BINDER**
	Headed by NT	**ADJT**
	Governed by SP/SPD/SPDA	
	- headed by NP(noun proper geographic)	**POBJ-LOC**
	- else	**POBJ**

Table 3: Role assignment rule table

And here below is the sentence we use to show the conversion process:

#ID=sent_01144							
0	restano	VIN(verb_intrans_tensed)		IBAR	-	CL(main)	
1	valide	AG(adjective)		ACOMP	0	SA	
2	le	ART(article)		SN	3	SN	
3	multe	N(noun)	S_TOP	0	SN		
4	già	AVV(adverb)		ADJM	3	SAVV	
5	irrogate	PPAS(past_participle_absolute)		MOD	3	SV3	
6	','	PUNT(sentence_internal)		SN	3	SN	
7	per	P(preposition)			ADJ	3	SP
8	le	ART(article)		SN	9	SN	
9	quali	REL(relative)		BINDER	7	SN	
10	pende	VIN(verb_intrans_tensed)		IBAR	3	IBAR	
11	il	ART(article)			SN	12	SN
12	giudizio	N(noun)	S_TOP	10	SN		
13	davanti_al	PHP(preposition_locution)	MOD	12	SP		
14	Tar	NPRO(noun_proper_institution)		POBJ	13	SN	
15	.	PUNTO(sentence_final)		F	0	F	

Table 4. Full conversion from phrase structure to dependency structure

As a final result, the treebank has 10,607 constituents with subject role, 3,423 of which have been manually assigned because they are in non-canonical position. Among the 7,184 SUBJ labels which were automatically identified, 46 constituents should have been assigned another function, with a precision of 0.99. On the other hand, 218 constituents should bear a SUBJ label instead of their actual label, with a recall of 0.97

4. A Quantitative Study of VIT

In this second part of the chapter, we introduce and discuss the quantitative data concerning the written portion of VIT and the constituents present in the 10.200 utterances of its Treebank. In particular, we will focus on some structures that are interesting from a parsing point of view and are called "stylistic" structures.

In a recent paper, Corazza et al. (2004) use a portion of VIT – 90.000 tokens produced in the SI-TAL project – to verify the possibility to train a statistic-probabilistic parser on the basis of procedures already experimented in English with PT by Collins and Bikel. Since the results they obtained are quite scarce (inferior to 70% accuracy), the authors wonder whether the poor performance might be due to intrinsic difficulties in the structure of the Italian language, to the different linguistic theory that has been adopted (cf. the lack of a VP node) or to the different tagset adopted, more detailed if compared to the one used in the PT.

According to what stated by Bikel regarding Collins' work, still a landmark for the creation of probabilistic parsers, the work done for the creation of a language model is to be anticipated by an important phase of preprocessing. This means that in order to produce the language model one does not work on the raw data of a treebank, but on a version modified on purpose. Collin's aim was to capture the biggest amount of regularities with the smallest number of parameters.

Probabilities are associated to lexicalized structural relations, i.e. structures where the head of the constituent to encode is present, that aim at helping to make decisions concerning the choice of arguments vs. adjuncts, of levels of attachment of a modifier and other similarly important matters otherwise difficult to capture when using only tags. For this purpose, it was necessary to intervene on the treebank by marking complements, sentences with null or inverse subject, and so on.

The preprocessing task accomplished by Corazza et al. is summarized here below and is actually restricted to the use of lemmas in place of word forms as head of lexicalized constituents:

> "As a starting point, we considered Model 2 of Collins' parser [7], as implemented by Dan Bikel [1], as its results on the WSJ are at the state-of-

> the-art. This model applies to lexicalized grammars approaches traditionally considered for probabilistic context-free grammars (PCFGs). Each parse tree is represented as the sequence of decisions corresponding to the head-centered, top-down derivation of the tree. Probabilities for each decision are conditioned on the lexical head. Adaptation of Collins' parser to Italian included the identification of rules for finding lexical heads in ISST data, the selection of a lower threshold for unknown words (as the amount of available data is much lower), and the use of lemmas instead of word forms (useful because Italian has a richer morphology than English; their use provides a non negligible improvement). At least at the beginning, we did not aim to introduce language-dependent adaptations. For this reason no tree transformation (analogous to the ones introduced by Collins for WSJ) has been applied to ISST."(p.4)

From the verifications carried out using two different parsers, researchers have come to the conclusion that,

> "These preliminary results... confirm that performance on Italian is substantially lower than on English. This result seems to suggest that the differences in performance between the English and Italian treebanks are independent of the adopted parser... our hypothesis is that the gap in performance between the two languages can be due to two different causes: intrinsic differences between the two languages or differences between the annotation policies adopted in the two treebanks."(p.5-6)

From the experiment computed on the basis of the information theory it turns out that the difference in performance cannot be imputed to the amount of rules and therefore to the type of annotation introduced, but to the scarce predictability of their structural relations, as stated by the authors,

> "First of all, it is interesting to note how the same coverage on rules results in the Italian corpus in a sensibly lower coverage on sentences (26.62% vs. 36.28%). This discrepancy suggests that missing rules are less concentrated in the same sentences, and that, in general, they tend to be less correlated the one with the other. This would not be contradicted by a lower entropy, as the entropy does not make any hypothesis on the correlation between rules, but only on the likelihood of the correct derivation. This could be a first aspect making the ISST task more difficult than the WSJ one. In fact, the choice of the rules to introduce at each step is easier if they are highly correlated with the ones already introduced."(p. 9)

4.1. Regularity and discontinuity in the language and its linguistic representation

A number of conclusions can be safely drawn from what the researchers stated and from the results of their test. Intuitively one could assert that the better the structural regularity of a language or its representation is, the wider its reproducibility on a statistical basis; on the contrary, in a language containing many cases recurring only once, in general hapax, bis-, tris-

legomena, a good statistical result of the model is less probable – this is called sparcity/sparseness. In linguistic terms the issue can be due to the division of grammar into core and periphery and this partition should be characterized in a quantitative manner. A statistical parser needs a great number of canonical structures belonging to the core grammar and it is not a case that in his procedure of creation of the model Collins deliberately introduces some corrections in the original treebank; that is, one has to accurately account for the structures which compose the core grammar, while the ones that constitute the periphery are amended ad hoc. Therefore, the malfunctioning of a statistical parser trained on a treebank must be related to the reference linguistic framework chosen by the annotators and hence to the reference language.

From the global quantitative data reported in Table 2. below, one can see that much more than half of the Italian sentences (9.800 in 19.099) do NOT have a subject lexically expressed in canonical position: this makes it very aleatory to locate the SN Subject. If we compare this with PT we get a completely different picture. For instance, in PT there are 4647 sentences which have been classified with the node of topicalized structure (S-TPC) which includes argument preposing, sentences in direct reported speech, and so on. Moreover there are sentences with an inverse structure, classified as SINV, only 827 of which are also TPC: SINV sentences are 2587 and they all typically have the subject in post-verbal position.

While as for the work on PT it is sensible to correct the problem in the pre-processing phases as made by Collins and commented by Bikel, in our case this issue is less sensible and certainly more complicated. In fact, the SN subject can be realized in four different ways: it can be lexically omitted, it can be found with an inverted position in the COMP constituents where complements are placed, it can be found in dislocated position on the left or on the right of the sentence to which it is related, at CP level. In a preliminary annotation of such cases we counted a total of more than 3000 cases of lexically expressed subject in non-canonical position. Then there are about 6000 cases of omitted subject to be taken into account. All these sentences must be dealt with in different ways during the creation of the model.

If one considers that in PT there are 93532 sentence structures – identifiable with the reg_ex "(S (" - 38600 of which are complex sentences, that is the 41% of all the "(S (" – adding up all the cases of non-canonical SUBJect sums up to a very low percentage, around 1%. On the contrary, in VIT the same phenomenon has a much higher percentage, over 27% in the case of non-canonical structures, and over 50% as to the omitted or unexpressed subject. We have also taken into consideration the annotation of complements in non-canonical position, and they have been listed in a table below.

Treebanks	Non-	Structures	Total (TU)	Total	Totale

Vs. Non-canonical Structures	canonical Structures (TU)	with Non-Canonical Subject (TS)	Utterances	(TS) Simple Sentences	Complex Sentences
VIT	3719	9800	10,200	19,099	6782
Percentage	27.43%	51.31%	63.75%		66.5%
PT	7234	2587	55,600	93,532	38,600
Percentage	13.01%	0.27%	59.44%		69.4%

Table 5. Comparison of non-canonical Structures in VIT and in PTB where we differentiate TU (total utterances) and TS(total simple sentences)

Here below in Table 6. we show absolute values for all non-canonical structure we relabeled in VIT. Considering that the total number of canonical lexically expressed SUBJects is 7172, we can compute the number of non-canonical subjects as constituting 1/3 of all expressed SUBJects – total number of lexically expressed subjects corresponding to 10,100. We labeled as S_TOP subject NPs positioned to the right of the governing verb; as S_DIS those subject NP which are positioned to the left of the governing verb but are separated from it by a parenthetical or a heavy complement; S_FOC are typically subject in inverted postverbal position of presentational structures; finally LDC are all types of Left Dislocated Complements with or without a doubling clitic.

LDC (left dislocated complements)	**S_DIS** (dislocated subject)	**S_TOP** (topicalized subject)	**S_FOC** (Focalized Subject)	**Total Non-Canonical**
251	**1037**	**2165**	**266**	**3719**

Table 6. Non-canonical Structures in VIT

5. Ambiguity and Discontinuity in VIT

We will briefly present and discuss some of the most interesting structures contained in VIT as regards the two important question of ambiguity and discontinuity in Italian. The most ambiguous structures are constituted by Adjectival related structures. As already commented above, adjectives in Italian may be positioned in front or after the noun they modify almost freely for most lexical classes. Only few classes require to be in predicative position and a very small number of adjectives must be placed in front of the noun they modify, in attributive position. A count of the functional conversion of adjectival structures is presented here below:

1296 Complement APs (ACOMP), 18748 Modifiers (MOD), 324 Adjuncts (ADJ), 2001 COORDinate APs

5.1. Ambiguous Predicative SA

Postnominal adjectives constitute the most challenging type since they may be considered as either post or premodifiers of a following nominal head. Even though postnominal non-adjacent SA recur in a small number – only 5.34%, they need to be identified by the parser. In the examples below we try to show how this process requires knowledge of adjectival lexical class besides feature matching. For every example taken from VIT we report the relevant portion of structure and a literal translation in a line below preceded by a slash.

(1) sn-[art-i, n-posti,
spd-[partd-della, sn-[n-dotazione, sa-[ag-organica_aggiuntiva]]],
sa-[ag-disponibili, sp-[p-a,
/the posts of the pool organic additive available to

Syntactic ambiguity arises and agreement checking is not enough even though in some cases it may solve the attachment preferences for the predicative vs. the attributive position.

(2) sn-[sa-[ag-significativi], n-ritardi]],
sn-[sa-[ag-profonde], n-trasformazioni],
ibar-[vt-investono],
/significative delays profound transformations affect

Adjectival structures may come in a row and modify different heads as in,

(3) sn-[art-il, n-totale,
spd-[partd-dei, sn-[n-posti,
spd-[partd-della, sn-[n-dotazione, sa-[ag-organica]]],
ag-vacanti], sa-[ag-disponibili
/the total of the posts of the pool organic additive vacant available

where "vacant" modifies the local head "posti", as well as "disponibili" which however governs some complement. On the contrary, in the example below, "maggiori" is not attached to the a possible previous head "orientamenti", but to a following one as the structure indicates,

(4) ibar-[vin-darebbe],
compin-[sp-[in-anche, part-agli,
sn-[n-orientamenti, spd-[pd-di, sn-[n-democrazia, sa-[ag-laica]]]],
sn-[sa-[ag-maggiori
/would give also to the viewpoints of democracy laic main

5.2 Sentence Complement

Another interesting phenomenon SA is their ability to head Sentential Complements: however in case of copulative constructions they are nominalized SA, as in the following example.

(5) f-[sn-[art-il,
sa-[ag-bello]],

ibar-[vc-è],
compc-[fac-[pk-che]
/the beatiful is that

5.3 Tough Problems: Quantification

As can be noted, in both cases complement sentences have an implicit impersonal subject pronoun. Structures which constitute tough problems to represent are Quantified structure. They can be SQ or SC, i.e. Quantifier Phrase or Comparative Phrase. Let's consider some example for both cases:

(6) sq-[in-molto, q-più, coord-[sa-[ag-efficace, punt-„, ag-controllabile, cong-e, ag-democratico]],
sc-[ccom-di,
f2-[sq-[relq-quanto],
cp-[savv-[avv-oggi],
f-[ibar-[neg-non, vcir-sia]
/much more effective , controllable and democratic of how much today not be

where we see a case of coordinate SA governed by the quantifier operator PIU'. Here is another case,

(7) cp-[sc-[ccom-tanto, sq-[q-più],
f-[ibar-[vc-sono], compc-[sa-[ag-lunghi]]],
sc-[ccom-tanto, sq-[q-maggiore],
f-[ibar-[vc-è],
compc-[sn-[art-la, n-soddisfazione, sa-[ag-finale]
/much more are long much higher is the satisfaction final

where we see cases of comparative structures at sentence level. On the contrary the following example is a case of quantification in the form of a relative construction,

(8) cp-[
cp-[sa-[ag-generali],
sp-[p-per, f2-[relq-quanto,
f-[ir_infl-[vcir-siano]]]], punt-„,
f-[sn-[art-le, n-regole], ibar-[vt-investono
/general for as much as be the rules involve

5.4 Fronted SPs in Participials

Another interesting construction present in Italian is the possibility to have fronted PP complements in Participials. This structure may cause ambiguity and problems of attachment, as shown in the examples below,

(9) sp-[p-in,
sn-[n-base, sp-[part-al,
sn-[n-punteggio,
sv3-[sp-[p-ad, sn-[pron-essi]],
ppas-attribuito, compin-[sp-[p-con,

/on the basis of the scoring to them attributed with

where we see that “ad essi” could be regarded as a modifier of the previous noun “punteggio”, whereas it is a complement of “attribuito” which however follows rather precede it.
Another more complex case is constituted by,

(10) sp-[p-a,
coord-[sn-[sa-[ag-singoli], n-plessi], cong-o, sn-[n-distretti],
sv3-[sp-[p-in, sn-[pron-essi]], ppas-compresi, punto-.]]]]]]]]]
/to single groups or districts in them comprised

The structure is not only found in bureaucratic language but also in literary genre, as in,

(11) spd-[partd-della,
sn-[n-cortesia,
sv3-[sp-[p-in, sq-[q-più, pd-di, sn-[art-un_, n-occasione]]],
vppt-dimostrata,
compin-[coord-[sp-[p-a, sn-[pron-me]],
/of the courtesy in more than one occasion demonstrated to me

5.5 Subject Inversion and Focus Fronted APs

Other non canonical structures are constituted by Subject Inversion, Focus Inverted APs, Left Clitic Dislocation with Resumptive pronoun.
A very frequent construction is constituted by the possibility to invert the Subject NP in postverbal position. This is usually linked to the presence of an Unaccusative verb governing the sentence.

(12) f-[ibar-[vc-diventa],
compc-[savv-[avv-così],
sa-[in-più, ag-acuta],
sn-[art-la, n-contraddizione], sp-[p-tra
/becomes so more acute the contradiction between

the same may happen with copulative verbs, where we see however that the subject is postponed after the open SA complement,

(13) f-ibar-[vc-è],
compc-[sa-[ag-peculiare,
sp-[part-all, sn-[np-Italia]]],
sn-[art-l, n-esistenza, spd-[pd-di
/is peculiar to Italy the existence of

Here is a cases of Fronted APs,

(14) cp-[s_foc-[ag-Buono],
f3-[sn-[cong-anche, art-l, n-andamento,
spd-[partd-delle, sn-[n-vendite

/good also the behaviour of the sales

All these structures are quite peculiar to the Italian language and also belong stylistically to a certain domain – financial news – and type of newspaper,

5.6 Hanging Topic and Left Clitic Dislocation

Italian allows to move locally in front of the utterance a portion of information which is somewhat resumed in the following sentence or may be left implicit and constitutes an elliptical material. Resumption usually takes place with a clitic pronoun. When the material fronted is not separated by a comma – a pause – it becomes a case of Left Clitic Dislocation.

(15) cp-[ldc-[art-una, n-decisione, sa-[ag-importante]],
f-[sn-[nh-Ghitti],
ibar-[clitac-l, ausa-ha, vppt-riservata],
/a decision important Ghitti it has reserved

(16) cp-[ldc-[sa-[ag-altra], n-fonte,
spd-[pd-di, sn-[n-finanziamento]]],
f-[ibar-[vc-sarà],
compc-[sn-[art-il, n-trattamento
/other source of funding will be the treatment

The one below is a case of Hanging Topic

(17) cp-[sn-[sa-[ag-brutta], n-faccenda], punt-„
f-[sn-[art-i, n-sudditi],
ibar-[clit-si, vt-ribellano, punto-.]]
/bad story , the populace self rebel

5.7 Aux-to-Comp Structures

Finally we will present and discuss some Aux-to-comp structures attested again both in bureaucratic and literary genres.

(18) cp-[f-[sn-[art-La, n-perdita],
sp-[p-per, sn-[art-il, npro-Rolo]],
ibar-[vcir-sarebbe],
compc-[congf-però,
spd-[pd-di, sn-[in-circa, num-'30', num-miliardi]]]],
topf-[auxtoc-[auag-avendo],
f-[sn-[art-la, npro-Holding],
sv3-[vppt-incassato,
compt-[sn-[n-indennizzi,
sp-[p-per, sn-[num-'28',
num-miliardi]]]]]], punto-.]
/the loss for the Rolo would be then of about 30 billion having the Holding cashed payments for 28 billions

Here the gerundive auxiliary precedes the subjec NP which in turn precedes the lexical verbal head in participial form. Below is a typical only Italian aux-to-comp structure,

(19) fc-[congf-e, punt-',',
topf-[auxtoc-[clit-si, aueir-fosse],
f-[sn-[pron-egli],
sv3-[vppin-trasferito, cong-pure,
compin-[sp-[part-nel,
sn-[sa-[in-più, ag-remoto], n-continente]]]]]
/and , self would be he moved also in the more remote continent

This case and the following only belong to literary genre,

(20) cp-[sn-[topf-[auxtoc-[art-l, ausai-avere],
f-[sn-[art-il, n-figlio],
sv3-[vppt-abbandonato,
compt-[sn-[art-il, n-mare],
sp-[p-per, sn-[art-la, n-città]]]]]],
f-[ibar-[clitdat-le, ause-era, avv-sempre, vppt-sembrato]
/the have the son abandoned the sea for the city her was always seemed

Peculiarities in common with classical aux-to-comp is the presence of an auxiliary as structural indicator of the beginning of the construction. We introduced a new special constituent TOPF to include the auxiliary and the sentence where the lexical verbal head has to be searched in order to produce an adequate semantic interpretation.

5.8 (In)Direct Reported Speech

Now we will present some cases of sentential structures which are/should be marked by special punctuation to indicate reported Direct or Indirect speech. In all these sentences we have treated the governing sentence which usually is marked off by commas or by dashes, as a parenthetical. We will briefly comment 4 types of constructions as follows,

- parenthetical inserted between SUBJ and IBAR
- parenthetical inserted between material in CP and the F
- free reported direct speech and then quoted direct speech
- Direct speech is ascribed to an anonymous "someone" quoted anyhow

(21) dirsp-[par-",
cp-[sp-[p-a, sn-[sa-[dim-questo], n-punto]],
f-[sn-[art-la, n-data],
par-",
fp-[punt-,, f-[ibar-[ausa-ha, vppt-detto],
compt-[sn-[npro-d_, npro-Alema],
savv-[avv-ieri], nt-sera]]], punt-,],
par-", ibar-[vin-dipende],
/" at this point the date " , said D'Alema last night , " depends

As can be noted, quotes individuate the portions of the utterance which is reported Direct Speech. The question is that the Subject NP “la data”/the date is separated from the Main Verb by the presence of the parenthetical governing clause. Below is a similar case,

(22) dirsp-[par-",
cp-[sp-[p-in, sn-[sa-[dim-questo], n-libro]],
f-[sn-[nh-madre, npro-Teresa],
fp-[par--, f-[ibar-[vt-spiegano],
compt-[sp-[part-alla,
sn-[npro-Mondadori]]]], par--],
ir_infl-[vcir-darà],
/“in this book Mother Theresa -- explain at the Mondadori - will give

Punctation does not help much in this example since the parenthetical is just introduced without indicating the end of the Reported Direct Speech.

5.9 Residual Problems: Relatives And Complement Clauses As Main Sentences

Italian allows freely to use Relative and Complement (with complementizer) Clauses as main clauses. This is due partly to the fact that Latin allowed it freely. But certainly it can be regarded a stylish way of organizing a text.

(23) cp-[f2-[rel-Che,
cp-[fp-[punt-,, f-[ibar-[vt-sostengono],
compt-[sp-[part-alla, sn-[npro-Farnesina]]]], punt-,],
f-[ibar-[neg-non, ausa-ha,
sp-[p-per, avvl-niente],
vppt-gradito],
compt-[sn-[art-l, n-operazione, n-by_pass]],
punto-.]]]]
/That , maintain at the Farnesina , not has in no case liked the operation by_pass .

This example has the additional problem of the presence of a parenthetical sentence which should indicate the presence of an Indirect Reported Speech structure. Certainly hard to spot.

(24) cp-[fac-[pk-che, savv-[avv-poi],
f-[sn-[art-la, n-legge],
ibar-[neg-non, virin-riesca],
compin-[sv2-[pt-a, viin-funzionare]]]],
punt-,,
f-[ibar-[vc-è],
compc-[sn-[art-un, n-discorso, f2-[rel-che
/That then the law not manages to work , is a matter that

6. Preliminary Evaluation

Here below we present preliminary data made available by Alberto Lavelli from IRST/ITC who implemented Bikel's model and parser on VIT with the usual machine learning procedure of 10-Fold Cross Validation. The first table refers to the homogeneous subset of VIT composed of sentences from the financial newspaper called "Il Sole-24 Ore".
Here below we present data related to the whole of VIT. As can be noticed, there is no remarkable difference.

Number of sentences = 10189
Number of Error sentences = 12
Number of Skip sentences = 0
Number of Valid sentences = 10177
Bracketing Recall = 68.61
Bracketing Precision = 68.29
Complete match = 8.70
Average crossing = 3.25
No crossing = 38.37
2 or less crossing = 61.73
Tagging accuracy = 96.65

Table 7a. Statistical parsing on complete VIT

Again a slight improvement is obtained when sentence length is reduced,

-- Sentence length<=40 --	
Number of sentences	**= 8519**
Number of Error sentences	**= 12**
Number of Skip sentences	**= 0**
Number of Valid sentences	**= 8507**
Bracketing Recall	**= 71.87**
Bracketing Precision	**= 71.58**
Complete match	**= 10.40**
Average crossing	**= 1.94**
No crossing	**= 45.47**
2 or less crossing	**= 71.72**
Tagging accuracy	**= 96.55**

Table 7b. Statistical parsing on complete VIT with sentence length limitation

VIT differs greatly from PT not only for the amount of sentences and data, but also for the choice to include linguistic material of different nature: in VIT there are five different genres – news, burocratic genre, political genre, scientific genre, literary genre -, while in PT only one is represented. Hence

the wider homogeneity we expect from PT and consequently the scarcer homogeneity in VIT.

The sparcity of VIT makes it difficult, if not impossible, to use it as a Language Model in the construction of probabilistic grammars for Italian. Therefore it is necessary to introduce corrective elements in order to enable the learning phase to distinguish sentences with different typologies (subject in canonical preverbal position, subject in non-canonical post-verbal position, lexically unexpressed subject, left dislocated "hanging Topic" subject – separated from the verb by other complements (or composed of a "heavy" SN followed by punctuation) - right dislocated Hanging Topic subject – separated from the verb by other complements), etc. . To this end, we implemented Bikel's language model directly on VIT and from preliminary results we can safely say that the same poor performance is reconfirmed – around 70% accuracy. More experiments will be carried out to confirm the hypothesis in Corazza et al., even though from the data in our possession such a confirmation is very likely.

References

Banfield A. 1982, *Un speakable Sentences. Narration and Representation in the Language of Fiction*, Boston , Routledge & Kegan Paul, pp.23-63.

Bianchi D., R.Delmonte(2002), Tecniche di apprendimento applicate al problema del tagging: una prima valutazione per l' Italiano, Workshop "Nlp e Web: La Sfida della Multimodalita' Tra Approcci Simbolici e Approcci Statistici", Convegno Nazionale AI*IA, Siena, pp.20-34.

Bikel, Daniel M. 2003. Intricacies of Collins' parsing model. Computational Linguistics, **30**(4), pp. 479-511.

Black, E. , S. Abney, D. Flickinger, C. Gdaniec, R. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini, and T. Strzalkowski: A Procedure for Quantitatively Comparing the Syntactic Coverage of English Grammars. In Proceedings of the DARPA Speech and Natural Language Workshop, pp. 306-311. 1991.

T. Brants. 2000. TnT: A statistical part-of-speech tagger. ANLP 2000.

E. Brill. 1995. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. Computational Linguistics 21: 543-565

Bristot A., Chiran L., R.Delmonte (2000), Verso una annotazione XML di dialoghi spontanei per l'analisi sintattico-semantica, XI Giornate di Studio GFS, Multimodalità e Multimedialità nella comunicazione, Padova.

Burchardt, A. , K. Erk, A. Frank, A. Kowalski, S. Pado and M. Pinkal: The SALSA Corpus: a German Corpus Resource for Lexical Semantics. In: Proceedings of LREC 2006, Genoa, Italy.

Carlson, Lynn, Daniel Marcu and Mary Ellen Okurowski: Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In: Jan van Kuppevelt and Ronnie Smith (eds.): Current Directions in Discourse and Dialogue. Kluwer Academic Publishers. 2003.

Carroll, J., T. Briscoe, A. Sanfilippo: Parser Evaluation: a Survey and a New Proposal. In Proceedings of the [First] International Conference on Language Resources and Evaluation, pp. 447-454. 1998.

Charniak E. 1997. Statistical Techniques for Natural Language Parsing. AI Magazine 18(4): 33-44.

Corazza A., Lavelli A., Satta G., Zanoli R. 2004. Analyzing an Italian Treebank with State-of-the-Art Statistical Parsers**.** In *Proceedings of the 3rd Workshop on Treebanks and Linguistic Theories (TLT-2004*), pp. 39-50, Tübingen, Germany.

Delmonte R.(1999), From Shallow Parsing to Functional Structure, in Atti del Workshop AI*IA -

"Elaborazione del Linguaggio e Riconoscimento del Parlato", IRST Trento, pp.8-19.
Delmonte R. 2000. Shallow Parsing And Functional Structure In Italian Corpora, LREC, Atene, pp.113-119.
Delmonte R.(2001), Tecniche e Strumenti per una Scomposizione e Rappresentazione Multilivello del contenuto linguistico di Dialoghi Spontanei (Tecniche e Strumenti per la rappresentazione di dialoghi), CNR, Roma.
Delmonte R. (2001), How to Annotate Linguistic Information in FILES and SCAT, in Atti del Workshop "La Treebank Sintattico-Semantica dell'Italiano di SI-TAL, Bari, pp.75-84.
Delmonte R., 2003 Multilevel linguistic transducers for the representation of spontaneous dialogues: from form to meaning in xml format, in "Voce, Canto, Parlato", Studi in Onore di Franco Ferrero, CNR, Padova, pp.117-134.
Delmonte R. 2003. Parsing Spontaneous Speech, in Proc. EUROSPEECH2003, Pallotta Vincenzo, Popescu-Belis Andrei, Rajman Martin "Robust Methods in Processing of Natural Language Dialogues" , Genève, pp, 1-6.
Delmonte, R. 2004. Strutture Sintattiche dall'Analisi Computazionale di Corpora di Italiano, in Anna Cardinaletti(a cura di), *Intorno all'Italiano Contemporaneo*, Franco Angeli, Milano, pp.187-220.
Delmonte R., Antonella Bristot, Luminita Chiran, Ciprian Bacalu, Sara Tonelli, Parsing the Oral Corpus AVIP/API, in Atti del Convegno Internationale "Il Parlato Italiano", Napoli, Universita' di Napoli, pp. 20
Delmonte R., Luminita Chiran, Ciprian Bacalu(2001), How to Integrate Linguistic Information in ***Files*** and Generate Feedback for Grammar Errors, Workshop on Sharing Tools and Resources for Research and Education, ACL, Toulouse, 10-14.
Delmonte R,.Luminita Chiran, Ciprian Bacalu, .(2000), Elementary Trees For Syntactic And Statistical Disambiguation, TAG+5, Paris, pp.237-240.
Delmonte R., R.Dolci(1989), Parsing Italian with a Context-Free Recognizer, **Annali di Ca' Foscari** XXVIII, 1-2, pp.123-161.
Delmonte R., E.Pianta, IMMORTALE - Analizzatore Morfologico, Tagger e Lemmatizzatore per l'Italiano, in Atti V Convegno AI*IA "Cibernetica e Machine Learning", Napoli, 19-22
Delmonte R., E.Pianta(1998), Immortal: How to Detect Misspelled from Unknown Words, in BULAG, PCUF, Besançon, 1998, 193-218.
Delmonte R., E.Pianta(1999), Tag Disambiguation in Italian, in Proc. Treebank Workshop ATALA, Paris, pp.43-49.
Michael Ellsworth, Katrin Erk, Paul Kingsbury and Sebastian Pado: PropBank, SALSA, and FrameNet: How Design Determines Product. In: Proceedings of the LREC 2004 Workshop on Building Lexical Resources from Semantically Annotated Corpora, Lisbon. 2004.
Gildea D. 2001. Corpus variation and parser performance. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing (EMNLP 2001)*, pp. 167–202, Pittsburgh, PA.
Esther König and Wolfgang Lezius: The TIGER language - A Description Language for Syntax Graphs. Part 1: User Guidelines. IMS, University of Stuttgart. 2002.
Marcu, Daniel, Magdalena Romera, and Estibaliz Amorrortu: Experiments in Constructing a Corpus of Discourse Trees: Problems, Annotation Choices, Issues. In: The Workshop on Levels of Representation in Discourse, pp 71-78, Edinburgh, Scotland, July 1999.
Montemagni et al.(2000), The Italian Syntactic-Semantic Treebank: Architecture, Annotation, Tools and Evaluation, LINC, ACL, Luxembourg, pp.18-27.
Montemagni, S. F. Barsotti, M. Battista, N. Calzolari, A. Lenci O. Corazzari, A. Zampolli, F. Fanciulli, M. Massetani, R. Basili R. Raffaelli, M.T. Pazienza, D. Saracino, F. Zanzotto, F. Pianesi N. Mana, and R. Delmonte 2003. Building the Italian Syntactic-Semantic Treebank. In Anne Abeillé, editor, *Building and Using syntactically annotated corpora*, pp. 189–210. Kluwer, Dordrecht.
Musillo G. & Khalil Sima'an 2002. Towards comparing parsers from different linguistic frameworks. An information theoretic approach. In *Proceedings of the LREC-2002 workshop Beyond PARSEVAL. Towards Improved Evaluation Measures for Parsing Systems*, pp. 44-51, Las Palmas, Spain.
Nivre, Joakim , Koenraad de Smedt and Martin Volk: Treebanking in Northern Europe: A White Paper. Nordisk Sprogteknologi. Nordic Language Technology. Årbog for Nordisk

Sprogteknologisk Forskningsprogram 2000-2004. Editor: Henrik Holmboe. Copenhagen. 2005.
Ulrik Petersen: Querying Both Parallel and Treebank Corpora: Evaluation of a Corpus Query System. Proc. of LREC. Genua. 2006.
Rizzi L., 1982. *Issues in Italian Syntax*, Foris Publications, Dordrecht.
Volk Martin and Yvonne Samuelsson: Bootstrapping Parallel Treebanks. In Proc. of Workshop on Linguistically Interpreted Corpora (LINC) at COLING. Geneva. 2004.
Volk, Martin, Sofia Gustafson-Capková, Joakim Lundborg, Torsten Marek, Yvonne Samuelsson and Frida Tidström: XML-based Phrase Alignment in Parallel Treebanks. In *Proc. of EACL Workshop on Multi-dimensional Markup in Natural Language Processing*. Trento. 2006.

APPENDIX I

1. Feature Structure or Dependency Representation

Parc 700 Dependency Bank
700 sentences from section 23 of the Upenn Wall Street Journal Treebank
http://www2.parc.com/isl/groups/nltt/fsbank/
Prague Arabic Dependency Treebank
100,000 words approximately
http://ufal.mff.cuni.cz/padt
Prague Dependency Treebank
1,5 million words
3 layers of annotation: morphological, syntactic, tectogrammatical
http://ufal.mff.cuni.cz/pdt2.0/
Danish Dependency Treebank
5,500 trees approximately
http://www.id.cbs.dk/~mtk/treebank/
Bosque, Floresta sinta(c)tica
10,000 trees approximately
http://acdc.linguateca.pt/treebank/info_floresta_English_html
French Functional Treebank
abeille@linguist.jussieu.fr
http://www.llf.cnrs.fr/Gens/Abeille/French-Treebank-fr.php
LinGO Redwoods
20,000 utterances (as for Fifth Growth)
http://lingo.stanford.edu/redwoods/
http://wiki.delph-in.net/moin/RedwoodsTop

2. Phrase Structure Representation

Penn Treebank
1 million words
dependency rules available for conversiion
http://www.cis.upenn.edu/~treebank/home.html
ICE – International Corpus of English
2million words tagged and parsed
http://www.ucl.ac.uk/english-usage/ice/
BulTreeBank
14,000 sentences
dependency version available
http://www.bultreebank.org/
Penn Chinese Treebank
40,000 sentences
http://www.cis.upenn.edu/~chinese/ctb.html
Sinica Treebank
61,000 sentences
http://godel.iis.sinica.edu.tw/CKIP/engversion/treebank.htm
Alpino Treebank for Dutch
150,000 words
http://www.let.rug.nl/vannoord/trees
TIGER/NEGRA
50,000/20,000 sentences
Dependency version available
http://www.ims.uni-struttgart.de/projekte/TIGER/TIGERCorpus
http://www.coli.uni-saarland.de/projekte/sfb378/negra-corpus/
TueBa-D/Z
22,000 sentences
Dependency version available
http://www.sfs.uni-tuebingen.de/en_tuebadz.shtml
TueBa-J/S
18,000 sentences
Dependency version available
http://www.sfs.uni-tuebingen.de/en_tuebajs.shtml
Cast3LB
18,000 sentences
Dependency version available
http://www.dlsi.ua.es/projectes/3lb/index_en.html
SUSANNE

Subset of Brown Corpus made up of 130,000 words
http://www.grsampson.net/Resources.html

3. Spoken Transcribed and Discourse Treebanks

Maptask
128 dialogues turned into 2597 files
there are similar efforts for other languages: Portuguese, Swedish, Dutch, Japanese
http://www.hcrc.ed.ac.uk/maptask/
PDTB – Penn Discourse TreeBank
Penn Treebank turned into Discourse Relation Treebank
http://www.seas.upenn.edu/~pdtb/
DGB – Discourse GraphBank
3110 sentences containing 8910 relations and clause pairs - 73Kwords
http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2005T08
RSTDT – Rhetorical Structure Theory Discourse Treebank
Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski
http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002T07
Talbanken05
300,000 words
http://w3.msi.vxu.se/~nivre/research/Talbanken05.html
Dependency version available
API-AVIP-IPAR - treebank
60,000 words - 5000 dialogue turns
http://www.cirass.unina.it/
CLIPS corpus
100 hours of spoken dialogues - phonetically annotated
http://www.clips.unina.it/
LIP corpus
500.000 tokens, 57 hours of spoken dialogues
fully tagged and lemmatized
http://languageserver.uni-graz.at/badip/badip/20_corpusLip.php
CHRISTINE
80500 words
http://www.grsampson.net/RChristine.html

4. Tools

@annotate
http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/annotate.html
Ananas
http://www.atilf.fr/ananas/
BulTreebank Project
http://www.bultreebank.org
CLaRK System
http://www.bultreebank.org/clark/
DTAG Treebank Tool
http://www.isv.cbs.dk/~mbk/dtag/
KPML development environment
http://www.fb10.uni-bremen.de/anglistik/langpro/kpml/README.html
LTChunk Systemic Coder
http://www.ltg.ed.ac.uk/~mikheev/tagger_demo.html
LBIS Coder
http://www.brain.riken.jp/labs/mns/sugimoto/LBISST/english.html
MMAX
http://www.eml-research.de/english/research/nlp/download/mmax.php
Poliqarp
http://poliqarp.sourceforge.net/
RST Tool for annotating with RST relations by Marcu
http://www.isi.edu/~marcu/software.html
SALSA
http://www.coli.uni-saarland.de/projects/salsa/
UAM Corpus Tool
http://www.wagsoft.com/CorpusTool/
SysFan tool
http://minerva.ling.mq.edu.au/
TnT tagger
http://www.coli.uni-saarland.de/~thorsten/tnt/
Wordfreak
http://wordfreak.sourceforge.net/
FreeLing
http://garraf.epsevg.upc.es/freeling/

5. Other resources based on treebanks

ACE project: PropBank/VerbNet/FrameNet
http://verbs.colorado.edu/~mpalmer/projects/ace.html
FrameNet
http://framenet.icsi.berkeley.edu/
NomBank

http://nlp.cs.nyu.edu/meyers/NomBank.html
NomLex
http://nlp.cs.nyu.edu/nomlex/index.html
ComLex
http://nlp.cs.nyu.edu/comlex/index.html

6. Generic website for corpora and other linguistic resources

http://www.corpus-linguistics.com/html/nav/main.html

http://www.ai.mit.edu/projects/iiip/nlp.html

http://billposer.org/Linguistics/Computation/Resources.html

http://nlp.stanford.edu/links/linguistics.html

http://www.bmanuel.org/

http://www.bmanuel.org/clr/clr2_tt.html

http://www.glue.umd.edu/~dlrg/clir/arabic.html

http://www.ims.uni-stuttgart.de/info/FTPServer.html

http://www.lai.com/mtct.html

http://www.aclweb.org/index.php?option=com_content&task=view&id=31&Itemid=31