Chapter 11

BUILDING THE ITALIAN SYNTACTIC-SEMANTIC TREEBANK

Simonetta Montemagni, Francesco Barsotti, Marco Battista, Nicoletta Calzolari, Ornella Corazzari, Alessandro Lenci, Antonio Zampolli
ILC-CNR / CPR, Pisa (Italy)

Francesca Fanciulli, Maria Massetani, Remo Raffaelli
Synthema, Pisa (Italy)

Roberto Basili, Maria Teresa Pazienza, Dario Saracino, Fabio Zanzotto
"Tor Vergata" University / CERTIA, Rome (Italy)

Nadia Mana, Fabio Pianesi
ITC-IRST, Trento (Italy)

Rodolfo Delmonte
Venice University / CVR, Venice (Italy)

Abstract The paper reports on the design and construction of a multi-layered corpus of Italian, annotated at the syntactic and lexico-semantic levels, whose development is supported by dedicated software augmented with an intelligent interface. The issue of evaluating this type of resource is also addressed.

Keywords: Italian, Multi-layered Linguistic Annotation, Syntactic Annotation, Semantic Annotation, Annotation Software, Evaluation

1. INTRODUCTION

Linguistically annotated corpora nowadays have a crucial theoretical as well as applied role in Natural Language Processing (NLP). This paper describes a large scale effort to provide Italian with a multi-level annotated corpus, the Italian Syntactic-Semantic Treebank (henceforth referred to as ISST). ISST – which represents one of the main activities of an ongoing Italian national project, SI-TAL[1] – was developed by a consortium of companies and computational linguistics sites in Italy which represent a wide range of expertise in the computational linguistics field[2].

Expected uses for ISST range from NLP tasks (such as Information Retrieval, Word Sense Disambiguation, linguistic knowledge acquisition) to training (and/or tuning) of grammars and sense disambiguation systems, to the evaluation of language technology systems.

The final resource consists of a multi-layered corpus of Italian of 305,547 word tokens. A first evaluation of the resource was carried out in the framework of an automatic Italian-English machine translation system. Both annotation and evaluation activities were supported by software, including an intelligent interface, specifically developed for efficiently accessing and querying the large amount of textual data and related annotations.

2. ISST ARCHITECTURE

ISST has a four-level structure covering morpho-syntactic, syntactic and lexico-semantic levels of linguistic description. Syntactic annotation is distributed over two different levels: the constituent structure level and the functional relation level. The fourth level deals with lexico-semantic annotation, which is carried out in terms of sense tagging augmented with other types of semantic information.

Both syntactic and lexico-semantic annotations refer to the morpho-syntactically annotated text, which in turn is linked to the orthographic file with the text and mark-up of macrotextual organisation (e.g. titles, subtitles, summary, body of article, paragraphs).

The multi-level structure of ISST shows two main innovations with respect to other treebanks. While most treebanks are restricted to syntactic annotation only, ISST – together with e.g. the Sinica (Chen et al., this volume) and the Prague Dependency (Böhmová et al., this volume) treebanks – includes both syntactic and semantic annotation levels. However, whereas the Chinese and the Prague dependency treebanks conceive of semantic annotation as the marking of thematic-like structures, ISST presents a further innovation given the fact that semantic annotation is intended here as sense tagging of lexical heads. In this way, the prerequisites are set up for corpus-based investigations on the syntax-semantics interface: the linking of the syntactic and semantic

annotation layers will permit, for instance, the identification of specific subcategorisation properties associated with a specific word sense, or of the semantic types associated with the functional positions of a given predicate.

The other innovative aspect of ISST concerns the distributed approach to syntactic annotation. In this respect, ISST differs from most treebanks, currently available or under construction for different languages, which adopt a unique syntactic representation layer, following either a constituency-based approach (see, among many others: Taylor et al., this volume; Marcus et al., 1993; Sampson, 1995, this volume; Greenbaum, 1996, Wallis, this volume; Moreno et al. 1999, this volume) or a dependency-based one (e.g. Karlsson et al., 1995), or a hybrid one combining features of both (e.g. Brants et al., this volume; Abeillé et al., this volume). ISST also differs from multi-level treebanks like the Prague Dependency Treebank (PTD). Whereas PTD annotation levels refer respectively to (a) the surface dependency relations and (b) the underlying sentence structure, ISST syntactic annotation levels are intended to provide orthogonal views of the same surface syntax.

3. ISST CORPUS

The ISST corpus consists of 305,547 word tokens reflecting contemporary language use. It includes two different sections:

1 a "balanced" corpus (215,606 tokens), exemplifying general language usage which consists of a selection of articles from newspapers (*La Repubblica* and *Il Corriere della Sera*) and periodicals, all selected to cover a high variety of topics (politics, economy, culture, science, health, sport, leisure, etc.);

2 a specialised corpus (89,941 tokens) from the financial domain, with articles taken from *Il Sole-24 Ore*.

All in all, they cover a 10 year time period (1985-1995).

4. ISST MORPHO-SYNTACTIC ANNOTATION

Morpho-syntactic annotation involves mark-up of morphological words with specification of part of speech, lemma, and morpho-syntactic features (such as number, person, gender, etc.). The adopted morpho-syntactic tagset conforms to the EAGLES international standard (Monachini and Calzolari, 1996) and includes 16 basic POS tags (see Appendix 1), which are further subdivided into 31 subtypes (to distinguish, for instance, between proper and common nouns, or demonstrative, possessive and clitic pronouns). The complete tagset, deriving from the combination of different POS subtypes with morpho-syntactic features, amounts to 236 different tags.

ISST morpho-syntactic annotation also includes treatment of (i) morphologically complex words such as cliticised verbs, e.g. *dammelo* lit. 'give+to_me+it', and (ii) basic multi-word expressions made up of contiguous sequences of words. The typology of multi-word expressions includes: expressions with words which do not occur separately, either foreign words (e.g. *prima_facie*) or Italian words which do not freely occur in texts (e.g. *chetichella* which only occurs in the adverbial locution *alla_chetichella* 'furtively'); expressions made up of sequences of tags not conforming to the general rules of grammar (e.g. *al_di_là* lit. 'at_of_there' meaning 'beyond' which is made up of two prepositions, the first of which also includes definiteness marking, followed by an adverb); expressions whose grammatical properties do not directly follow from their component words (e.g. multi-word prepositions such as *in_funzione_di* lit. 'in_function_of' meaning 'as'). Note, however, that in ISST other types of multi-word expressions are identified and marked at higher annotation levels (see section 6).

Morpho-syntactic annotation was previously carried out at ILC in the framework of the European projects PAROLE (Goggi et al., 1997) and ELSNET (Monachini and Corazzari, 1995). The text was automatically tagged by a stochastic tagger, the Pi-Tagger (Picchi, 1994); the output was then manually revised by a team of linguists.

5. ISST SYNTACTIC ANNOTATION

As already pointed out in section 2, ISST follows a distributed approach to syntactic annotation based on a monostratal view of syntax. ISST syntactic annotation levels are independent of each other: none of them presupposes the other, e.g. functional annotation is not built on top of constituent structure annotation. This makes it possible for the two annotation levels to be used (i.e. accessed and examined) independently. At the same time, they provide complementary information: in principle, combined views of the developed resource can be obtained, for example, by projecting functional information onto the constituent structure (see Montemagni et al., 2000).

The motivations for this "double track" approach to syntactic representation range from language-specific to more general.

This distributed approach to syntax is particularly suited to deal with some peculiarities of Italian syntax, namely:

- the syntactically free constituent order, which allows for considerable variation in the ordering of constituents at the sentence level;
- the pro-drop property: the subject of main verbs appears to be omitted in approximately 70% of the possible cases (Bates, 1976).

These two features together make a pure constituency-based representation of Italian unrestricted texts difficult: if on the one hand they can be handled in the annotation through the use of empty elements (either traces or pro-subjects), on the other hand their frequency of occurrence would make syntactic annotation less transparent and thus less intelligible. This is a well-known problem of syntactic annotation of free word order languages, see for instance (Brants et al., this volume).

ISST overcomes the problems of a pure constituency-based representation of Italian by decoupling functional information from the constituent structure. Hence, in ISST the treatment of word order variation does not interfere in any way with the representation of functional relations, i.e. the encoding of the latter becomes entirely separate from the order of constituents in the sentence. By the same token, subject omission is not accounted for at the level of constituent structure, but only at the functional level.

The distributed approach to syntactic annotation is also supported by the types of uses foreseen for ISST. In fact, ISST syntactic annotation levels are intended to be usable both in real applications and for research purposes, and to be compatible with different approaches to syntax, both dependency- and constituency-based, adopted in both theoretical and applied frameworks.

5.1 Constituency annotation

ISST constituency annotation departs from other constituency-based syntactic annotation schemes (e.g. the one adopted in the Penn Treebank) in a number of respects, mainly due to the distributed organisation of syntactic annotation.

Annotation at this level consists of the identification of phrase boundaries with labelling of constituent types. Given that functional relations are handled at a distinct level, ISST tree structures are shallow, as exemplified below for the sentence *lo scontro sulle cessioni legali è stato risolto per decreto* 'the clash on legal transfers has been resolved by decree':

```
(1) [F [SN lo scontro [SP sulle [SN cessioni [SA legali SA] SN] SP] SN]
    [IBAR e stato risolto IBAR] [COMPT [SP per [SN decreto SN] SP]
    COMPT] F]
```

In the constituent structure above, all the complements of the verb – the subject nominal constituent (SN), the verbal node (IBAR) and the complements node (COMPT) – are at the same level of embedding with respect to the sentence node (F). Similar observations hold for the internal structure of nominal constituents, where no hierarchical distinction is made among the head, the determiner and the complements and/or adjuncts (see the internal structure of the subject noun phrase in (1) above). Note also that, for verbal phrases, annotation is restricted to the minimal verbal nucleus (auxiliaries, negation, verb and clitics of inherently pronominal verbs), because the traditional notion

of VP (which includes the verb complements) is not easily applicable to unrestricted Italian texts given the problem of discontinuity, which is so frequent that it becomes controversial whether the notion of VP is really useful for the purpose of corpus annotation.

Moreover, the fact that in ISST functional relations are dealt with at a distinct level instead of being defined in terms of constituent structures allows ISST to dispense with empty elements such as null subjects or traces, thus making constituent annotation more intelligible. In fact, the relevant information is recovered at the functional level, through a relation linking the displaced or omitted element to its governing head. Therefore, syntactic phenomena such as pro-drop and ellipsis, as well as cases of discontinuous or non canonical order of constituents (topicalisation, wh-questions, etc.) are not accounted for in terms of empty categories and coindexation as e.g. in the Penn Treebank but rather at the functional annotation level. Examples of constituency-based representations of these structures follow:

Ho cose più importanti di cui occuparmi '(I) have more important things to take care of'

```
(2) [F [IBAR Ho IBAR] [COMPC [SN cose [SA piu importanti SA] [F2 [SPD
    di cui [SV2 occuparmi SV2] SPD] F2] SN] COMPC] F]
```

Gli ordini di vendita stranieri hanno imboccato la strada che riporta al di là del confine 'the foreign selling orders took the way which goes back beyond the border'

```
(3) [F [SN Gli ordini [SPD di [SN vendita SN] [SA stranieri SA] SPD]
    SN] [IBAR hanno imboccato IBAR] [COMPT [SN la strada [F2 che [F
    [IBAR riporta IBAR] [COMPIN [SP al_di_la_del [SN confine SN] SP]
    COMPIN] F] F2] SN] COMPT] F]
```

Constituency annotation in ISST uses an inventory of 22 constituent types (see Appendix 2): specialised constituent names are used for a number of complements or adjuncts, in order to facilitate the mapping onto functional annotation.

Constituency annotation of ISST was performed in a semi-automatic way. First, the text was parsed by a Shallow Parser (Delmonte, 1999, 2000) whose task was that of building shallow syntactic structures for each safely recognisable constituent. In uncertain cases, no attachment was performed at this stage in order to avoid committing to structural decisions which might then reveal themselves to be wrong. The output of the shallow parser was then manually revised and corrected.

5.2 Functional annotation

Functional annotation in ISST is word-based, i.e. it is carried out independently of previous identification of phrasal constituents. The advantages of this choice include, on the theoretical front, the fact that ISST can be used as a reference resource for a wider variety of different annotation schemes, both constituency- and dependency-based ones (Lin, 1998). Moreover, word-based functional annotation is comparatively easy and "fair" for use in parsing evaluation since it overcomes some of the well-known shortcomings of constituency-based evaluation (see, among others: Carroll et al., 1998, this volume; Lin, 1998, this volume; Sampson, 2000).

We used FAME (Lenci et al., 1999, 2000) as the starting point for the development of the ISST functional annotation scheme[3]. FAME's main features are: (a) the hierarchical organisation of functional relations which makes provision for underspecified representations of highly ambiguous functional analyses, and (b) the modular coding architecture which is articulated over different information layers, each factoring out different but possibly interrelated linguistic facets of syntactic annotation. These features combined make FAME a meta-scheme, i.e. an annotation scheme which, beyond being a full-fledged annotation scheme, also acts as a sort of "metalanguage" for different annotation schemata.

The building blocks of FAME are functional relations which are expressed in terms of binary relations holding between two lexical heads. Note that FAME relations involve words belonging to major lexical classes only (i.e. non-auxiliary verbs, nouns, adjectives and adverbs); information about grammatical words (e.g. determiners, prepositions, auxiliaries) is encoded otherwise (see below).

Functional relations include dependency, i.e. head-dependent, relations such as subject and object, which – unlike constituency annotation – can also involve displaced elements, null subjects and elliptical material; the inventory of ISST dependency relations is given in Appendix 3. This dependency-based annotation scheme is augmented with other relation types dealing with constructions which cannot be interpreted in terms of head-dependent relationships, e.g. coordination phenomena, clause-internal co-reference etc. For the sake of paper length, only dependency relations are discussed below. A dependency relation is an asymmetric binary relation between two full words, a head and a dependent. Each dependency relation is modularly represented as follows:

```
(4) dep_type (head.<head_features>,
              dependent.<dep_features>)
```

where dep_type specifies the relationship linking the dependent to the head and features associated with the elements of the relation further specify relational

information. Consider below the functional representation of the sentence *lo scontro sulle cessioni legali è stato risolto per decreto* 'the clash on legal transfers has been resolved by decree' (whose constituent structure representation is reported in (1) above):

```
(5)  sogg (risolvere.<diatesi=passiva>, scontro)
     mod (scontro, cessione.<intro=''su''>)
     mod (cessione, legale)
     mod (risolvere.<diatesi=passiva>, decreto.<intro=''per''>)
```

It can be observed that features convey, for instance, information about the preposition which introduces the dependent in a given relation (see the INTRO attribute), or about the diathesis of the verbal head. Other information types conveyed by features concern the open/closed predicative function of clausal dependents (in this way control information is also encoded), the semantic role of modifiers (e.g. temporal, locative), etc.

Unlike constituency annotation, at this level either the head or the dependent can correspond to elliptical material; this makes it possible to represent pro-drop phenomena:

Ho cose più importanti di cui occuparmi '(I) have more important things to take care of'

```
(6)  sogg (avere, .<pers=1, numb=sing>)
```

Note that this modular representation, distributed over relations and features, provides the prerequisites for ISST to be used as a reference annotation scheme which is compatible with a wide range of theories and thus mappable onto different syntactic representation formats (for more details on the inter-translatability of FAME into other syntactic representation formats see Lenci et al. 1999, 2000).

Annotation at the functional level was carried out manually.

6. ISST LEXICO-SEMANTIC ANNOTATION

The methodology for annotation at this level takes advantage of two previous experiments in semantic tagging carried out at ILC in the framework of the SENSEVAL initiative (Calzolari et al., 2000) and of the ELSNET resources task group activity (Corazzari et al., 2000).

The following inventory of annotation units is identified and distinguished in ISST:

- USS: sense units corresponding to single lexical items;
- USC: semantically complex units expressed in terms of multi-word expressions (e.g. compounds, support verb constructions, idioms);

- UST: title sense units corresponding to titles (of newspapers, books, shows, etc.).

Unlike USS and USC which are annotated at one level only, titles receive a two-level annotation: at the level of individual component words and as a single title unit. Note that in ISST lexico-semantic annotation is restricted to content words, in particular to nouns, verbs and adjectives and corresponding multi-word expressions.

Lexico-semantic annotation consists of the assignment of semantic tags, expressed in terms of attribute-value pairs. ISST semantic tags convey three different types of information:

1 sense of the target word(s) in the specific context: ItalWordNet (henceforth, IWN) is the reference lexical resource used for the sense tagging task (CPR et al., 2000). IWN, developed from the EuroWordNet lexicon (Alonge et al., 1998), includes a general part and a specialised one with financial and computational terminology;

2 other types of lexico-semantic information complementing IWN sense assignments, e.g. marking figurative usages, idiomatic expressions, evaluative suffixation, neologisms, proper nouns, etc.;

3 information about the tagging operation, mainly notes by the human annotator used to ease and speed up the annotation process and its revision. The human annotator can keep track of problematic cases (e.g. cases of indistinguishable IWN senses, of ambiguous corpus contexts, etc.). Input of this type may also be useful for discussion with the team of IWN lexicographers with a view to prospective revisions and updating of the lexical resource.

Note that through the taxonomical organisation of IWN word senses an implicit assignment is made to the semantic types of the IWN ontology. In this way, ISST sense tagging can also be seen as semantic tagging.

As for the annotation methodology for this level, in order to ensure the consistent tagging of polysemous words and USC, the annotation was manually performed 'per lemma' and not sequentially, that is, word by word following the text.

Annotation strategy. Each annotation unit is tagged with the relevant sense according to IWN sense distinctions. When more than one IWN sense applies to the context being tagged, arbitrary sense assignments are avoided by resorting to underspecified annotations (expressed, for instance, in terms of disjunction over different IWN senses).

Sense assignment is augmented through specification of lexico-semantic tags conveying information not explicitly included in the reference lexical resource or further specifying it. These additional semantic tags are used to mark:

- a US or USC used in a figurative sense (either metaphoric or methonymic): e.g. the metaphoric use of *fulmine* in *essere un fulmine* lit. 'to be a lightning bolt' meaning 'to be fast like a lightning bolt';
- a US semantically modified through evaluative suffixation (e.g. *porticciolo* 'small port', *borsone* 'a large bag');
- the semantic type (e.g. human entity, artifact, institution, location) of proper nouns, either US (e.g. *IBM* tagged as a 'group') or USC (e.g. *Nuova Zelanda* 'New Zealand' which can be assigned the type of 'place');
- the USC subtype, e.g. compound (e.g. *certificato di credito del tesoro* 'treasury certificate'), idiom (e.g. *essere la chiave di volta* 'to be the keystone'), support verb construction (e.g. *entrare in vigore* 'to come into effect');
- the UST subtype, e.g. title of an opera (e.g. *La Boheme*) or of a newspaper (e.g. *Il Corriere della Sera*).

Sense assignments combined together with additional lexico-semantic information make the ISST annotated corpus more than a mere list of instantiations of the senses attested in the reference lexical resource. This annotation strategy makes the annotated corpus a repository of interesting lexico-semantic information, especially concerning lexico-semantic facts which are excluded – either programmatically or just by chance – from the reference lexical resource.

Let us consider the case of non-lexicalised uses. Consider the following contexts, where the target word – marked in bold – is used metaphorically:

(7) *La nuova* ***arma*** *di vendetta è l'indifferenza*
'the new weapon of revenge is indifference'
(8) *Gli argentini ricominciano a mancare* ***appuntamenti*** *con la storia*
'Argentinians start to miss appointments with the history again'

The metaphoric use of target words is specified through a specific tag (FIGURATO=metaf). As to sense assignment, *arma* in (7) is assigned the appropriate figurative sense (IWN sense 2); by contrast, *appuntamenti* in (8), representing an instance of non-lexicalised metaphor, is linked to the literal sense. Through the interaction between sense and feature information lexicalised and non-lexicalised figurative uses can be singled out in ISST: namely,

non-lexicalised metaphors are always linked to the literal sense. Similar observations hold in the case of semantic modification conveyed through evaluative suffixation: non-lexicalised cases are linked to the relevant sense of the stem word.

Feature assignment is also used to further specify sense distinctions which are left underspecified at the level of the reference lexical resource. Let us take the case of regular polysemies. Geographical proper nouns in the IWN lexicon are assigned a unique sense covering both the readings of 'group of people' and 'place'. Whenever possible, the annotator disambiguates between the two readings through the assignment of a specific feature as shown in the examples below:

(9) *La **Francia** si è sentita isolata*
'France felt isolated'
(10) *Perturbazione in arrivo dalla **Francia***
'Disturbance coming from France'

In both (9) and (10) *Francia* is assigned the underspecified IWN sense (sense 1); the two occurrences are then differentiated through the value assigned to the feature PROPER_NOUN which in (9) is assigned the 'group of people' value and in (10) the 'place' one.

It may also be the case that corpus annotation identifies multi-word expressions that are not recognised as such in the reference lexical resource, but behave as semantically complex units for the purposes of corpus annotation. This is the case of expressions such as *anni Sessanta* 'the sixties' which, being fully compositional and productive, do not appear as independent entries in the lexical resource. In this case, the annotator marks *anni Sessanta* as a USC with no specific sense assigned to it.

Corpus annotation can also shed light on the variability of multi-word expressions; in fact, multi-word expressions, going from compounds to support verb constructions and idiomatic expressions, when effectively used are prone to massive variation (Sinclair, 1996). To this end, semantically complex units, while being recorded in relation to a single lemma, are annotated as covering also modifiers which may optionally appear in the expression. Consider the example in (11) below where identified USC's correspond to the words marked in bold:

(11) ***tagliare le ali** a qn* 'to clip somebody's wings'
***tagliare le ultime ali** a un paese* 'to clip the last wings of a country'

The grey areas spotted by these few examples in which corpus annotation either diverges from the lexical resource or further specifies it can be seen – in perspective – as the starting point for revisions and refinements of both the

annotated corpus and the reference lexical resource. In this way, the annotated corpus proves to be a flexible resource, which is to some extent independent from the specific internal architecture of the lexicon selected as the reference resource. On the other hand, this type of corpus annotation can help to enrich or simply tune the reference lexical resource through the addition of missing entries (or simply variants) and senses.

7. THE MULTI-LEVEL LINGUISTIC ANNOTATION TOOL

The labour-intensive annotation task requires devoted tools to access efficiently the large amount of textual data and related annotations. From this perspective, both a data model and effective graphical representations are necessary. The annotation tool of ISST, GesTALt, features specific data models and graphical representations defined to comply with the different needs of the three levels of annotation. Building upon these data models, level-oriented subsystems are provided. The tool is also designed to ease the control of intra-level and inter-level coherence.

7.1 The linguistic data base

The ISST linguistic data model has been represented in the object-oriented formalism, which was selected for its flexibility. Defined data are directly used in the object-oriented database underlying GesTALt. For each level of annotation, a specific container was defined. The system (and its subsystems) manages a collection of documents, the corpus: this relation is represented in a class hierarchy. Moreover, the different level interpretations associated with sentences in the corpus are modelled respectively via the class of objects. To give the reader a flavor of the object modelling of linguistic structures, we present here the hierarchy describing constituency annotation (i.e. the class *synt_int*).

Constituency annotation is based on tree structures where both internal nodes and leaves are constituents (*const*). Leaves are called *basic constituents* (*b_const*), while internal nodes *complex constituents* (*c_const*). The resulting *synt_int sub-hierarchy* is depicted in Figure 11.1.

Complex constituents are collections of constituents, either basic or complex ones. A constituency-based syntactic interpretation is thus the complex constituent representing the interpretation of the whole sentence. This notion is modelled by the relation between the *c_const* class and the *synt_int* class in the hierarchy.

7.2 The visual representation of annotations

Managing the annotation of large sentences is cumbersome. Effective graphical representations are needed by both the annotator and the user to

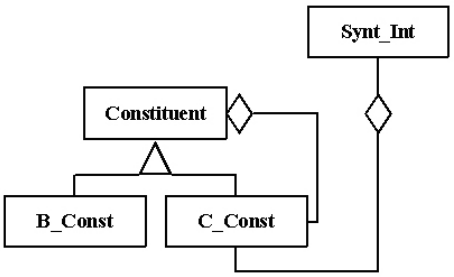

Figure 11.1. Syntactic interpretation

ease navigation in a complex information network. In what follows, the visual representation of syntactic and lexico-semantic annotations adopted within GesTALt is described.

Constituency-based Annotation. Constituency-based annotation is represented in terms of tree structures. A related manageable representation is thus needed within the annotation framework. Graphical tree representations can ease user interactions with the tree structures, i.e. retrieval and modification of encoded information.

Syntactic annotation of unrestricted texts involves long sentences (of 20 words or more). Sentence length (which determines the number of leaves) combined with the need for showing node tags may burden the tree diagram representation.

The visual representation defined for this annotation level is a *strip tree*, namely a tree described in strips (see Figure 11.2), which is similar to a bracketed representation but gives a hierarchical view of the structured information.

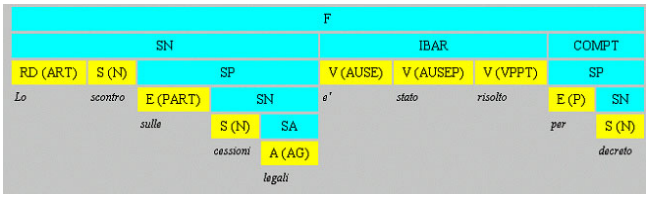

Figure 11.2. Strip tree

The annotation task at this level requires a convenient way of following the evolution of the tree structure. From this perspective, partial annotations have

to be represented, and the transition from one partial annotation to another has to be supported.

Functional annotation. Dependency-based functional annotations are visualised in terms of graphs. Participants are represented as nodes and functional relations as arcs. Hence, the subsystem devoted to functional annotation has to manage mainly insertion/deletion of functional relations (i.e. arcs) connecting nodes. Given that ellipsis phenomena are accounted for at this annotation level (see section 5), another important functionality of this annotation module involves the insertion/deletion of nodes corresponding to elliptical material.

An example of functional annotation (sentence (2) above), as it appears in the graphical interface, is given in Figure 11.3:

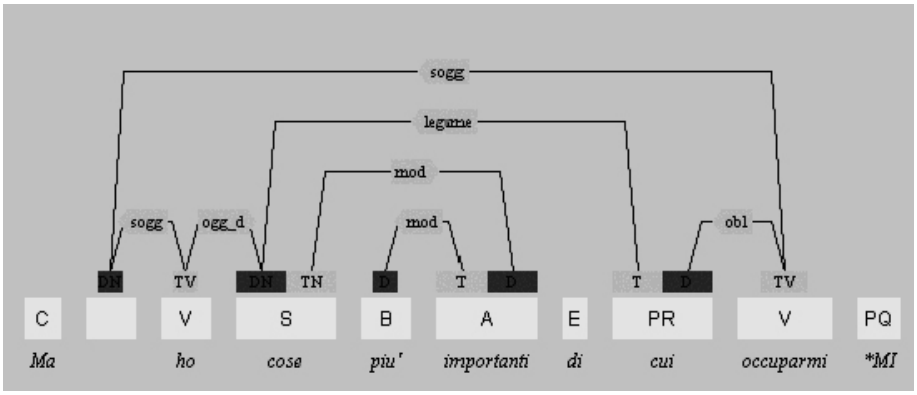

Figure 11.3. Planar graph

Functional relations are represented as directed labelled arcs linking words belonging to major lexical classes only; note that information about grammatical words is encoded in the feature structure describing each element of the relation. In the graphical interface, elliptical material is represented in terms of empty boxes: see the null subject in Figure 11.3. The example above also shows the representation adopted for morphologically complex forms: the word form *occuparmi* lit. 'occupy+me' is segmented into two different morphological words, corresponding to the verb and the clitic pronoun; annotation operates in fact on morphological words (see section 2).

Lexico-semantic annotation. The purpose of lexico-semantic annotation is to associate semantic tags to content words (namely, verbs, nouns and adjectives) and corresponding multi-word expressions. The graphical representation of lexico-semantic annotation is depicted in Figure 11.4.

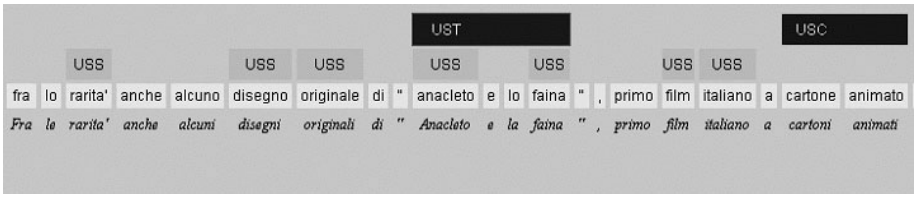

Figure 11.4. The visual representation of lexico-semantic annotation

This figure shows a semantically tagged sentence where different types of semantic units can be seen. Boxes labelled as USS represent semantic units corresponding to individual words; the USC box marks the identified semantically complex unit, corresponding to the compound noun *cartone animato* 'cartoon', and the UST box identifies a semantic unit of type title. In the case of semantically complex units, annotation – which is performed at one level only – involves sequences of words which are not necessarily contiguous. As for title units, annotation is carried out both at the level of the component words and of the title unit. Note that, for each identified semantic unit, annotation is represented in terms of a feature structure specifying the types of information described in section 6 above.

7.3 GesTALt architecture

The GesTALt annotation workbench is the resulting system, made up of a pool of cooperating subsystems. The system manages the linguistic database sketched in section 7.1 and produces its output in standard XML.

The system is a suite consisting of specific applications: SinTAS for constituency annotation; FunTAS for functional annotation; SemTAS for lexico-semantic annotation; and ValTAS for evaluation and correction of inter- and intra-level annotations.

FunTAS, SinTAS, and SemTAS are stand-alone applications. The synthesis of the three subsystems is achieved in ValTAS, which needs the capabilities spread throughout the subsystems. The technology adopted for the development (object-oriented design), in conjunction with an *ad hoc* architectural design, allows easy reuse of the functionalities developed for the subsystems in the global (i.e. ValTAS) system.

The overall GesTALt architecture is shown in Figure 11.5, where the components are represented as boxes, and the interactions as arrows.

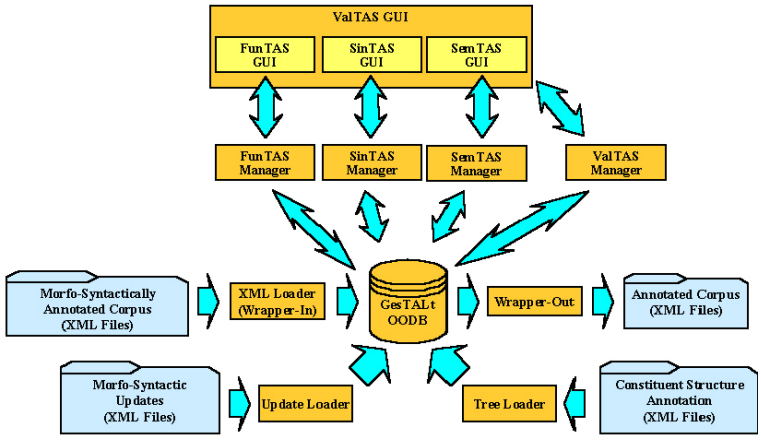

Figure 11.5. The visual representation of lexico-semantic annotation

The creating/translating flow of the object-oriented database (GestTALt-OODB) is shared by the subsystems. Information is downloaded from and uploaded to XML containers via specific wrappers (Wrapper-in and Wrapper-out). The GestTALt-OODB is the object-oriented database where the annotation of the different levels is stored respectively by FunTAS, SinTAS and SemTAS, together with the morphologically annotated corpus used as input by all annotation modules. Each subsystem, except for ValTAS which includes everything, is made up of specialised components. The graphical user interfaces based on the specific representations are depicted in the general architecture (FunTAS GUI, SinTAS GUI, SemTAS GUI and ValTAS GUI). Furthermore, the different ways of interacting with the database impose the design of special modules devoted to ad hoc navigation of the hierarchy (FunTAS Manager, SinTAS Manager, SemTAS Manager, and ValTAS Manager).

8. ISST EVALUATION

Information stored in ISST, particularly in the financial corpus, was used to improve an automatic Italian-English translation system, PeTra Word 2.0®, developed by Synthema and already on the market. The translation system is based on the Logical Grammars ("Slot Grammars") formalism (McCord 1980, 1989) and consists of three main components: the Italian language analyser (morphological analyser, monolingual dictionary and syntac-

tic parser), the transfer component (bilingual dictionary and structural transfer rules) and the English generator. Improvements were mainly concerned with mono- and bi-lingual dictionaries, the Italian grammar and the transfer rules.

Dictionaries were revised on the basis of corpus evidence. In particular, dictionary coverage was enlarged by adding missing specialised entries and by improving already existing ones; associated translations were also added to the bilingual dictionary. A crucial aspect of this enrichment was concerned with multi-word expressions: multi-word expressions annotated in ISST at the different levels were evaluated and eventually added to PeTra's dictionaries according to the system constraints, in terms of either individual entries or particular constructions associated with component words. This addition was essential for the analysis of multi-word adverbs or prepositions which could not be parsed otherwise (such as *al di là di*) and for the recognition of compound words which cannot be literally translated (such as *codice fiscale* 'taxpayer's code number'). ISST lexico-semantic annotation was also used to revise and improve the structure of the hierarchical semantic dictionary used e.g. for lexical transfer disambiguation.

Let us turn now to improvements at the level of analysis and transfer rules. Before the tuning of the system with ISST, the grammar already had a good coverage (i.e. 88% on unrestricted texts). In spite of this fact, there were syntactic constructions attested in the ISST corpus which were analysed incompletely or incorrectly: this also follows from the fact that the subcorpus selected for evaluation is a specialised one, containing syntactic structures not currently used in standard Italian. ISST was first examined to check grammar coverage: accessing ISST on the basis of functional relations which correspond to the slots, it was possible to study the features associated with them and their constituency-based representation. In this way, the main features of an uncovered syntactic structure were identified and included both in the grammar rules and at the level of syntactic transfer.

The adopted evaluation methodology can be summarised as follows: identification of a thematically homogeneous subcorpus; translation of the subcorpus before and after tuning; classification of translated sentences into correct translation, inaccurate translation (requiring minor revisions), wrong translation, sentence which could not be translated; analysis and comparison of the results obtained before and after the system tuning. It came out that correctness of results increased by 17%. By comparing the classification results in the two translation runs (before and after tuning) it was observed that the number of correctly translated sentences increased by 45%, and the number of inaccurate translations by 40%. As a consequence, the number of wrong translations decreased by 38%, and of untranslated sentences by 79%. This overall improvement of translation results led to a significant reduction (about 18%) of the time required for the manual revision of the translations.

9. CONCLUSION

The final tested version of ISST is now available together with the annotation and browsing software specifically developed within the project. ISST consists of 89,941 word tokens annotated at the constituency structure level, 305,547 at the functional level and 81,236 content words at the lexico-semantic level (corresponding to 69,972 identified semantic units). For about one third of the ISST corpus (namely the financial part) there are four annotation layers available simultaneously. The annotated corpus is also available in XML format. An overall description of achieved results can be found in (Montemagni and Pazienza, 2001).

Completion of course refers to the goals set up within the SI-TAL project, since resources like ISST require continuous refinement and extension. In fact, treebanks and NLP resources by their very nature should be regarded as open-ended enterprises. Two possible extensions might be envisaged, both along the horizontal dimension and along the vertical one. As for the former, ISST coverage can obviously be extended by adding new annotated texts, as well as spoken data and domain-specific corpora. On the other hand, vertical extensions might involve the enrichment of information encoded for existing annotation levels or – most importantly – the addition of new annotation layers (e.g. annotation of anaphoric relations or of discourse structure).

Notes

1. SI-TAL was a joint enterprise working towards an integrated suite of tools and resources for Italian Natural Language Processing, funded by the Italian Ministry of Science and Research (MURST) and coordinated by the Consorzio Pisa Ricerche (CPR).

2. Different partners were in charge of different aspects of the project: ILC-CNR/CPR, Venice University/CVR and ITC-IRST were in charge of the annotation, "Tor Vergata" University/CERTIA of the design and construction of the annotation software and Synthema of the evaluation of the developed resource.

3. FAME originated as a revision of a de facto standard, i.e. the functional annotation scheme developed in the framework of the LE-2111 SPARKLE project (Carroll et al., 1996), a revision first carried out to comply better with the basic requirements of parsing evaluation (in the framework of the LE4-8340 ELSE project), and then to make the scheme suitable for annotation of unrestricted Italian texts.

References

Abeillé A., Clément L., Toussenel F., (2003). Building a treebank for French. This volume.

Alonge A., Calzolari N., Vossen P., Bloksma L., Castellon I., Martì T., Peters W., (1998). The Linguistic Design of the EuroWordNet Database. *Computers and the Humanities* 32, 2-3, p. 91-115.

Bates E., (1976). *Language and Context: Studies in the Acquisition of Pragmatics*, New York, Academic Press.

Brants T., Skut W., Uszkoreit H., (2003). Syntactic annotation of a German newspaper corpus. This volume.

Böhmová A., Hajic J., Hladká B., Panenová J., (2003). The Prague Dependency Treebank: a three-level annotation scenario. This volume.

Calzolari N., Corazzari O., (2000). Senseval/Romanseval: the framework for Italian. *Computers and the Humanities*, vol. 34, n. 1-2, p. 61-78.

Carroll J., Briscoe T., Calzolari N., Federici S., Montemagni S., Pirrelli V., Grefenstette G., Sanfilippo A., Carroll G., Rooth M., (1996). *Specification of Phrasal Parsing*, SPARKLE Deliverable 1.1.

Carroll J., Briscoe E., Sanfilippo A., (1998). Parser Evaluation: a Survey and a New Proposal. *Proceedings of LREC*, Granada, p. 447-454.

Caroll J., Minnen G., Briscoe T., (2003). Parser evaluation using a grammatical relation annotation scheme. This volume.

Chen K., Luo C., Gao Z., Chang M., Chen F., Chen C., (2003). Sinica Treebank. This volume.

Corazzari O., Calzolari N., Zampolli A., (2000). An Experiment of Lexical-Semantic Tagging of an Italian Corpus. *LREC-2000 Proceedings*, Athens.

Corazzari O., Monachini M., (1995). *ELSNET: Italian Corpus Sample*, ILC, Pisa.

CPR, ITC-IRST, Quinary, (2000). *ItalWordNet: Rete semantico-lessicale per l'italiano*. SI-TAL, Specifiche Tecniche di SI-TAL, Manuale Operativo, Capitolo 2.

Delmonte R., (1999). From Shallow Parsing to Functional Structure. *Atti del Workshop AI*IA* "Elaborazione del Linguaggio e Riconoscimento del Parlato", IRST Trento, p. 8-19.

Delmonte R., (2000). Shallow Parsing and Functional Structure in Italian Corpora. *LREC-2000 Proceedings*, Athens.

Goggi S., Biagini L., Picchi E., Bindi R., Rossi S., Marinelli R., (1997). *Italian Corpus Documentation*, LE-PAROLE WP2.11, ILC, Pisa.

Greenbaum S. (ed.), (1996). *English Worldwide: The International Corpus of English*, Oxford, Clarendon Press.

Karlsson F., Voutilainen A., Heikkila J., Anttila A. (eds.), (1995). *Constraint Grammar, a language-independent system for parsing unconstrained text*. Berlin and New York, Mouton de Gruyter.

Lenci A., Montemagni S., Pirrelli V., Soria C., (1999). FAME: a Functional Annotation Meta-scheme for Multimodal and Multi-lingual Parsing Evaluation. *Proceeding of the ACL99 Workshop on Computer-Mediated Language Assessment and Evaluation in NLP*, University of Maryland.

Lenci A., Montemagni S., Pirrelli V., Soria C., (2000). Where opposites meet. A Syntactic Meta-scheme for Corpus Annotation and Parsing Evaluation. *LREC-2000 Proceedings*, Athens.

Lin D., (1998). A dependency-based method for evaluating broad-coverage parsers. *Natural Language Engineering*, 4(2), p. 97-114.

Lin D., (2003). Dependency-based evaluation of MINIPAR. This volume.

Marcus M., Marcinkiewicz M.A., Santorini B., (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2), p. 313-330.

McCord M.C., (1980). Slot Grammars. *Computational Linguistics*, vol. 6, p. 31-43.

McCord M.C., (1989). Design of LMT: A Prolog-based Machine Translation System. *Computational Linguistics*, vol 15, p. 33-52.

Monachini M., Calzolari N., (1996). *Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora. A Common Proposal and Application to European Languages*. EAGLES Recommendations. Pisa, ILC.

Montemagni S., Barsotti F., Battista M., Calzolari N., Corazzari O., Zampolli A., Fanciulli F., Massetani M., Raffaelli R., Basili R., Pazienza M.T., Saracino D., Zanzotto F., Mana N., Pianesi F., Delmonte R., (2000). The Italian Syntactic- Semantic Treebank: Architecture, Annotation, Tools and Evaluation. *Proceedings of the COLING Workshop on "Linguistically Interpreted Corpora (LINC-2000)*", Luxembourg, 6 August 2000, p. 18-27.

Montemagni S., and Pazienza M.T. (eds), (2001) *Atti del Workshop su "La Treebank sintattico-semantica dell'italiano in SI-TAL"*. 7° Congresso della Associazione Italiana per l'Intelligenza Artificiale (AI*IA 2001), Bari, 26 settembre 2001.

Moreno A., López A., Sánchez F., (1999). *Spanish Tree Bank: Specifications*, Version 4, Manuscript.

Moreno A., López A., Sánchez F., Grishman R., (2003). Developing a syntactically annotation scheme and tools for a Spanish treebank. This volume.

Picchi E., (1994). Pi-Tagger: A tagger and lemmatizer for Italian. *EURALEX-94 Proceedings*, Amsterdam.

Sampson G., (1995). *English for the Computer*, Oxford, Clarendon Press.

Sampson G., (2000). A proposal for improving the measurement of parse accuracy. *International Journal of Corpus Linguistics*, vol. 5, n. 1, p. 53-68.

Sampson G., (2003). Thoughts on two decades of drawing trees. This volume.

Sinclair J., (1996). The Empty Lexicon. *International Journal of Corpus Linguistics*, 1, p. 99-119.

SI-TAL, (2000). Specifiche Tecniche di SI-TAL. Manuale Operativo. ILC, Pisa.

Taylor A., et al. (2003). The Penn Treebank: an overview. This volume.

Wallis S., (2003). Completing parsed corpora from correction to evolution. This volume.

Appendix

1. Morpho-syntactic tagset

Tag	Description
S	Nouns
V	Verbs
A	Adjectives
P	Pronouns
T	Predeterminers
D	Determiners
R	Articles
B	Adverbs
E	Prepositions
C	Conjunctions
N	Numerals
I	Interjections
@@	Punctuation
SA	Abbreviations
X	Residuals

2. Syntactic constituents tagset

Constituent type	Description
F	Sentence
SN	noun phrase, including its complements and/or adjuncts
SA	adjectival phrase, including its complements and/or adjuncts
SP	prepositional phrase
SPD	prepositional phrase *di* 'of'
SPDA	prepositional phrase *da* 'by, from'
SAVV	adverbial phrase, including its complements and/or adjuncts
IBAR	verbal nucleus with finite tense and all adjoined elements like clitics, adverbs and negation
SV2	infinitival clause
SV3	participial clause
SV5	gerundive clause
FAC	sentential complement
FC	coordinate sentence (also ellipsed and gapped)
FS	subordinate sentence
FINT	+wh interrogative sentence
FP	punctuation marked, parenthetical or appositional sentence
F2	relative clause

Constituent type	Description
CP	dislocated or fronted sentential adjuncts
COORD	coordination with coordinating conjunction as head
COMPT	transitive/passive/ergative/reflexive complement
COMPIN	intransitive/unaccusative complement
COMPC	copulative/predicative complement

3. Hierarchy of dependency relations