The Italian Syntactic-Semantic Treebank: Architecture, Annotation, Tools and Evaluation

1] S. Montemagni, F. Barsotti, M. Battista, N. Calzolari, O. Corazzari, A. Zampolli
ILC-CNR / CPR, Pisa (Italy)
2] F. Fanciulli, M. Massetani, R. Raffaelli
Synthema, Pisa (Italy)
3] R. Basili, M.T. Pazienza, D. Saracino, F. Zanzotto
"Tor Vergata" University / CERTIA, Roma (Italy)
4] N. Mana, F. Pianesi
ITC-IRST, Trento (Italy)
5] R. Delmonte
Venezia University / CVR, Venezia (Italy)

Abstract

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

Introduction

It is nowadays widely acknowledged that linguistically annotated corpora have a crucial theoretical as well as applicative role in Natural Language Processing. Italian still lacks such a resource. The 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). Evaluation of ISST is foreseen in the framework of a machine translation application. Specifically developed software, including an intelligent interface, supports both annotation and evaluation activities.

ISST - which represents one of the main actions of an ongoing Italian national project, SI-TAL[1] - is developed by a consortium of companies and computational linguistics sites in Italy (see author's affiliations above). 1], 4] and 5] are in charge of the annotation, 3] of the design and construction of the annotation software and 2] of the evaluation of the developed resource.

Expected uses for ISST range from Natural Language Processing 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. ISST also promises to contribute to the start up of commercial systems for Italian processing. Last but not least, although annotated corpora are typically built and used in research and applicative contexts, their potential for teaching purposes has also to be emphasised; see, for instance, their use in the classroom for teaching syntax at Nijmegen University (Van Halteren 1997).

The final and tested version of ISST will be available in year 2001. Currently, the annotation phase is started, based on the linguistic guidelines and the annotation software which have just been released; yet, initial specifications remain subject to extensions and further refinements on the basis of feedback coming from the annotation process (e.g. emergence from the corpus of linguistic phenomena not yet covered by the specifications).

1 Architecture of ISST

ISST has a three-level structure ranging over syntactic and semantic levels of linguistic description. Syntactic annotation is distributed over two different levels, namely the constituent structure level and the functional relations level: constituent structure is annotated in terms of phrase structure trees reflecting the ordered

[1] SI-TAL is a joint enterprise leading 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).

arrangement of words and phrases within the sentence, whereas functional annotation provides a characterisation of the sentence in terms of grammatical functions (i.e. subject, object, etc.). The third level deals with lexico-semantic annotation, which is carried out here in terms of sense tagging augmented with other types of semantic information. The three annotation levels are independent of each other, and all refer to the same input, namely a morpho-syntactically annotated (i.e. pos-tagged) text which 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 final resource will be available in XML coding.

The multi-level structure of ISST shows two main novelties with respect to other treebanks:

- it combines within the same resource syntactic and lexico-semantic annotations, thus creating the prerequisites for corpus-based investigations on the syntax-semantics interface (e.g. on the semantic types associated with functional positions of a given predicate, or on specific subcategorisation properties associated with a specific word sense);
- it adopts a distributed approach to syntactic annotation which presents several advantages with respect both to the representation of the syntactic properties of a language like Italian (e.g. its highly free constituent order) and to the compatibility with a wide range of approaches to syntax.

2 ISST input

2.1 Corpus composition

ISST corpus consists of about 300,000 word tokens reflecting contemporary language use. It includes two different sections: 1) a "balanced" corpus, testifying general language usage, for a total of about 210,000 tokens; 2) a specialised corpus, amounting to 90,000 tokens, with texts belonging to the financial domain.

The balanced corpus contains a selection of articles from different types of Italian texts, namely newspapers (*La Repubblica* and *Il Corriere della Sera*) and a number of different periodicals which were selected to cover a high variety of topics (politics, economy, culture, science, health, sport, leisure, etc.). The financial corpus includes articles taken from *Il Sole-24 Ore*. All in all, they cover a 10 year time period (1985-1995).

2.2 Morpho-syntactic annotation

Syntactic and lexico-semantic annotation takes as input the morpho-syntactically annotated text. 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; the output was manually revised by a team of linguists. The adopted morpho-syntactic tagset conforms to the EAGLES international standard (Monachini and Calzolari 1996).

Annotation at this level involves identification of morphological words with specification of part of speech, lemma, and morho-syntactic features such as number, person, gender, etc.

Morphological words typically stand in a one-to-one relation with orthographic words with two exceptions, namely: i) the case of more than one morphological word which forms part of the same orthographic word (as in the case of cliticized words, e.g. *dammelo* 'give+to_me+it'); ii) the case of more than one orthographic word which make up a single morphological word not otherwise decomposable (as in the case of multi-word expressions such as *ad_hoc*, *al_di_là* 'beyond', *fino_a* 'up_to').

3 ISST annotation schemata

3.1 General requirements

The design of each individual annotation schema underlying ISST and their interrelations are intended to fit a list of basic requirements following directly from the typology of foreseen uses. They include:

a) usability in both real applications and research purposes;
b) compatibility with different approaches to syntax, both dependency- and constituency-based, either adopted in theoretical or applicative frameworks;
c) applicability on a wide scale, in a coherent and replicable way;
d) applicability to both written and spoken language (this requirement does not apply to the actual ISST but it is foreseen in view of

possible resource extensions to spoken language data).

Within ISST, requirements a) and b) are satisfied by distributing the annotation over different levels (mainly for what concerns syntactic annotation) and, for each level, by factoring out different information types according to different dimensions.

Different strategies are pursued to meet requirement c). This is achieved at the level of individual annotation schemes by first providing wide coverage and detailed annotation criteria and then by avoiding as much as possible arbitrary annotation decisions (i.e. uncertainty cases are preferably dealt with through underspecification or disjunction over different interpretations). c) has also consequences on the relationship between different annotation levels: redundancy is avoided as much as possible; i.e. a given information type has to be specified only once, at the relevant annotation level (e.g. grammatical relations such as subject and object are only specified at the functional level).

Finally, d) is guaranteed by the independence of syntactic annotation levels: spoken data, which are typically fraught with ellipses, anacolutha, syntactic incompleteness and other related disfluency phenomena cannot be easily represented in terms of constituency. By contrast, the level of functional analysis - which in ISST has an independent status - naturally reflects a somewhat standardised representation, since it abstracts away from the surface realisation of syntactic units in a sentence, thus being relatively independent of disfluency phenomena and incomplete phrases.

3.2 Syntactic annotation

Most treebanks, currently available or under construction for different languages, adopt a unique syntactic representation layer, following either a costituency-based approach (see, among many others, Marcus et al. 1993, Sampson 1995, Greenbaum 1996, Sandoval et al. 1999) or a dependency-based one (e.g. Karlsson et al. 1995), or a hybrid one combining features of both (e.g. Brants et al. 1999, Abeillé et al. 2000). ISST departs from all of them since it adopts a multi-level structure.

To our knowledge, the only multi-level treebank is the Prague Dependency Treebank (PTD, Bémová et al. 1999), but in this case the different annotation levels refer respectively to a) the surface dependency relations and b) the underlying sentence structure. By contrast, ISST adopts a monostratal view of syntax, and thus both syntactic annotation levels are rather intended to provide orthogonal views of the same surface syntax. These views, though complementary, are developed independently of each other.

This bi-level approach to syntactic representation is particularly suited to deal with a language like Italian, which allows for considerable variation in the ordering of constituents at the sentence level. In fact, by decoupling functional information from the constituent structure, 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 contituents in the sentence.

3.2.1 Constituency annotation

In 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, due to: a) the peculiarities of Italian as a free constituent order language; b) the distributed organisation of syntactic annotation in ISST.

Constituency annotation in ISST uses an inventory of 22 constituent types (see table below). Specialized constituent names are used for a number of complements or adjuncts, in order to help the mapping with functional annotation.

Const type	**Meaning**	**Classif**
F	sentence	structural
SN	noun phrase, including its complements and/or adjuncts	substantial
SA	adjectival phrase, including its complements and/or adjuncts	lexical
SP	prepositional phrase	lexical
SPD	prepositional phrase *di* 'of'	lexical
SPDA	prepositional phrase *da* 'by, from'	lexical
SAVV	adverbial phrase, including its complements and/or adjuncts	substantial
IBAR	verbal nucleus with finite tense and all adjoined elements like clitics, adverbs and negation	substantial
SV2	infinitival clause	substantial
SV3	participial clause	substantial
SV5	gerundive clause	substantial

Const type	Meaning	Classif
FAC	sentential complement	lexical
FC	Coordinate sentence (also ellipsed and gapped)	lexical
FS	Subordinate sentence	lexical
FINT	+wh interrogative sentence	lexical
FP	punctuation marked, parenthetical or appositional sentence	lexical
F2	relative clause	lexical
CP	dislocated or fronted sentential adjuncts	structural
COORD	Coordination with coordinating conjunction as head	lexical
COMPT	Transitive/Passive/Ergative/Reflexive Complement	structural
COMPIN	Intransitive/Unaccusative Complement	structural
COMPC	Copulative/Predicative Complement	structural

From the point of view of their relations to functional labelling, syntactic constituents are divided up into two main subgroups (see column 3 of the table above): functional constituents and substantial constituents. This subdivision reflects theoretical assumptions which are derived from the Lexical Functional Grammar theory. In particular, functional constituents are internally subdivided into structural constituents (used to set complements apart) and lexical constituents (headed by a lexical head with or without semantic content). Structural constituents also contain F and CP where F has the task of indicating the canonical sentential constituent and CP indicates the presence of sentential adjuncts, or some discontinuity in the utterance.

At the same time, 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 annotation more intelligible. In fact, the relevant information is recovered at the functional level, through a relation linking the displaced element to its head. Therefore, syntactic phenomena such as pro-drop, 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.

Constituency annotation of ISST is worked out in a semi-automatic way. First, the text is parsed by a Shallow Parser (Delmonte 1999, 2000) whose task is that of building shallow syntactic structures for each safely recognizable constituent. In uncertainty cases, no attachment is performed at this stage in order to avoid being committed to structural decisions which might then reveal themselves to be wrong. In fact, it is preferable to perform some readjustment operations after structures have been built rather than introducing errors from the start. Then, the output of the shallow parser is manually revised and corrected.

3.2.2 Functional annotation

Functional annotation in ISST is carried out by marking relations between words belonging to major lexical classes only (i.e. non-auxiliary verbs, nouns, adjectives and adverbs), independently of previous identification of phrasal constituents. 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, on the applicative side, head-based functional annotation is comparatively easy and "fair" to be used for parsing evaluation since it overcomes some of the well-known shortcomings of constituency-based evaluation (see, among others, Carroll et al. 1998, Sampson 1998, Lin 1998). Last but not least, head-based functional annotation is naturally i) multi-lingual, as functional relations probably represent the most significant level of syntactic analysis at which cross-language comparability makes sense, and ii) multi-modal, since it permits comparable annotation of both spoken and written language.

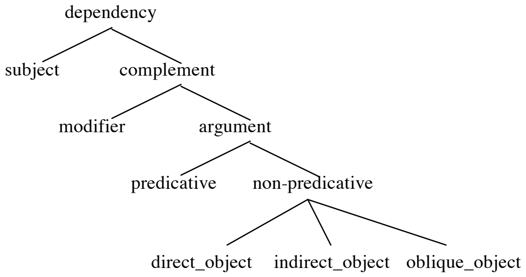
FAME (Lenci et al. 1999, 2000) is the annotation scheme (originally developed in the SPARKLE project LE-2111 and then revised in the framework of ELSE LE4-8340) adopted for functional annotation in ISST, which has been revised and integrated to make it suitable for annotation of unrestricted Italian texts. The building blocks of FAME are functional relations, further subdivided into dependency relations and other relation types dealing with coordination phenomena and clause-internal co-referential bonds. Only the former are described below for sake of paper length.

A dependency relation is an asymmetric binary relation between full words, respectively a head and a dependent. Each dependency relation is expressed as follows:

```
dep_type (lex_head.<head_features>,
          dependent.<dep_features>)
```

where dep_type specifies the relationship holding between the lexical head and its dependent. At this level, either the head or the dependent can correspond to elliptical material; this makes it possible to deal with pro-drop phenomena and other types of elliptical constructions.

Dep_types are hierarchically structured to make provision for underspecified representations of highly ambiguous functional analyses (see above). The typology of dependency relations, hierarchically organized, is given below.



In the proposed scheme a crucial role is played by the features associated with the elements of the relation, which complement relational information. Features convey, for instance, information about the grammatical word (preposition or conjunction) which possibly introduces the dependent in a given relation, or about the open/closed predicative function of clausal dependents (in this way control information is also encoded).

Functional annotation in ISST is thus modularly represented, i.e. it is structured into relational and feature information, each factoring out different but interrelated facets of functional annotation. This modular representation 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 intertranslatability of FAME into other syntactic representation formats see Lenci et al. 1999, 2000).

Annotation at the functional level is carried out manually.

3.2.3 Annotation examples

The sketchy description of the syntactic annotation schemes provided above is complemented here with annotation examples. The two ISST syntactic annotation levels, the constituent structure and the functional ones, are developed independently; in spite of this fact, they are strictly interrelated and complement each other.

In order to show the peculiarities of the two annotation levels and their interrelations, let us consider the ISST annotation of the following Italian sentence, *Giovanni sembra arrivare domani* 'John seems to arrive tomorrow':

- **Constituent structure annotation**

```
f-[    sn-[Giovanni],
       ibar-[sembra],
       sv2-[arrivare,
              savv-[domani]]]
```

- **Functional annotation**

```
sogg (sembrare, Giovanni)
arg (sembrare,
       arrivare.<status= aperto>)
mod (arrivare, domani)
sogg (arrivare, Giovanni)
```

Note that the subject relation holding between *arrivare* and *Giovanni* in the functional annotation does not find an explicit counterpart at the level of constituent structure representation since subject raising is not treated at that level.

Depending on the expected uses, the two annotation layers can be accessed and examined independently. However, due to the complementarity of the information contained in them, combined views on the developed resource can also be obtained. For instance, projection of functional information onto the constituent structure results as follows:

```
f-[    sn-sogg[Giovanni],
       ibar-[sembra],
       sv2-arg[arrivare,
                 savv-mod[domani]]]
```

where each constituent category is marked, whenever possible, with a functional tag. This is one of the many possible combined views which can be obtained on the ISST syntactically annotated corpus.

3.3 Lexico-semantic annotation

3.3.1 Basics

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

In ISST, lexico-semantic annotation consists in the assignment of semantic tags, expressed in terms of attribute/value pairs, to full words or sequences of words corresponding to a single unit of sense (e.g. compounds, idioms). In particular, annotation is restricted to nouns, verbs and adjectives and corresponding multi-word expressions.

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 two parts, a general one and a specialized one with financial and computational terminology;
2) other types of lexico-semantic information not included in the reference lexical resource, e.g. for marking of figurative uses;
3) information about the tagging operation, mainly notes by the human annotator about problematic annotation cases.

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.

Starting from the assumption that senses do not always correspond to single lexical items, the following typology of annotation units is identified and distinguished in ISST:

us: sense units corresponding to single lexical items (either nouns, verbs or adjectives);

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 any type (of newspapers, books, shows, etc.). Titles receive a two-level annotation: at the level of individual components and as a single title unit.

3.3.2 Annotation criteria

Each annotation unit is tagged with the relevant sense according to IWN sense distinctions. In order to meet requirement c) in section 3.1 above, arbitrary sense assignments, which may occur when more than one IWN sense applies to the context being tagged, are avoided by means of underspecification (expressed in terms of disjunction/conjunction over different IWN senses).

The other lexico-semantic tags allow to mark:

- a us or usc used in a metaphoric or methonymic or more generally in a figurative sense: e.g. *la molla di una simile violenza* 'the spring of such a violence' where *molla* is used in a metaphoric sense. The distinction between lexicalized and non lexicalized figurative usages can be inferred from the assigned IWN sense: non lexicalized figurative uses are linked to the literal sense;
- a us semantically modified through evaluative suffixation (e.g. *appartamentino* 'small flat', *concertone* 'big concert');
- the semantic type (i.e. human entity, artifact, institution, location, etc.) of proper nouns, either us (e.g. *pds* 'the pds party' is semantically tagged as a 'group') or usc (e.g. *Corno d'Africa* 'the Horn of Africa' is assigned the sematic type of 'place');
- the usc subtype, e.g. compound (e.g. *prestito obbligazionario* 'loan stock'), idiom (e.g. *mettere i puntini sulle i* 'to dot one's i's'), support verb construction (e.g. *dare aiuto* 'to give assistance');
- the ust subtype, i.e. title of an opera (e.g. *Il barbiere di Siviglia*), of a newspaper (e.g. *La Nazione*) or of something else.

In this way, the annotated corpus provides more than a list of instantiations of the senses attested in the reference lexical resource. Through the added value of this additional information, the annotated corpus becomes a repository of interesting semantic information going from titles and proper nouns to non-lexicalized metaphors, metonymies and evaluative suffixation, and in general to non-conventional uses of a word.

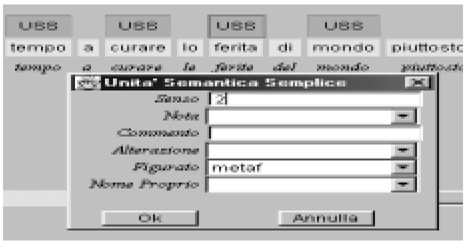
Finally, notes about the tagging operation are mainly used to ease and speed up the annotation process and its revision: the human annotator can keep track of problematic cases (e.g. cases of indistinguishable IWN senses, of ambiguous corpus contexts, etc.). Input of this type may also be useful for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

As to the annotation methodology for this level, in order to ensure that polysemous words and usc are tagged consistently, the annotation is manually performed ‘per lemma’ and not sequentially, that is, word by word following the text.

3.3.3 Annotation examples

Let us exemplify the annotation strategy illustrated in the previous sections with a few semantically tagged corpus occurrences.

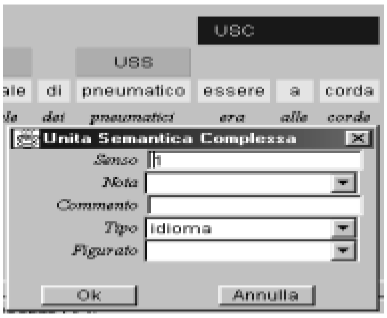
An example of an annotated us is given below: the target word is *ferite* 'wounds' in the context *curare le ferite del mondo* 'to cure the wounds of the world'. In the annotation window, the target word is assigned the sense number 2; the feature *figurato=metaf* marks its metaphoric use in the specific corpus context.



Unita' Semantica Semplice

Senso: 2
Nota:
Commento:
Alterazione:
Figurato: metaf
Nome Proprio:

Annotation of semantically complex units (usc) is exemplified below for the multi-word expression *essere alle corde* 'to be hard-pressed':

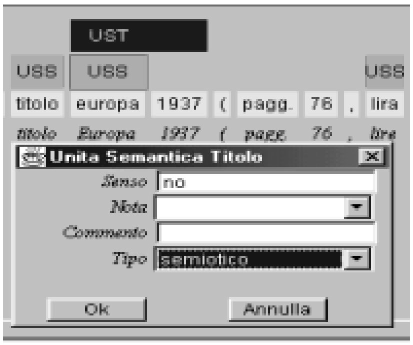


Unita Semantica Complessa

Senso: 1
Nota:
Commento:
Tipo: idioma
Figurato:

The blue box covering the text shows that it has been marked as a usc; the annotation window specifies its sense number (1) in IWN and its type (idiom).

Finally, an example is given below for title sense units, or ust. It can be noticed that the book title *Europa 1937* 'Europe 1937' is annotated both at the level of its constituting words (see *Europa*) and as a single unit of type title of a book (*tipo=semiotico*). Obviously, sense information does not apply to ust.



Unita Semantica Titolo

Senso: no
Nota:
Commento:
Tipo: semiotico

4 The multi-level Linguistic Annotation Tool

The labour intensive annotation task demands for tools devoted to access efficiently the large amount of textual data and the related annotations. In this perspective, both a data model and effective graphical representations are mandatory.

GesTALt is the annotation tool defined for ISST where an object oriented data definition has been preferred for its flexibility. Specific data models and graphical representations are defined so to comply with the different needs of the three levels of annotation. Building upon these data models, level-oriented subsystems are settled. The tool is also designed to ease the control of intra-level and inter-levels coherence.

4.1 The linguistic data base

The model of linguistic data is designed within the object oriented formalism. The defined data are directly used in the object oriented database underlying *GesTALt*. For each level of annotation, a specific container has to be 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 modeled respectively via the class of objects. To give the flavor of the object modeling of linguistic structures, we present here the hierarchy describing constituent 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 resultant *synt_int* sub-hierarchy is depicted in Fig. 1.

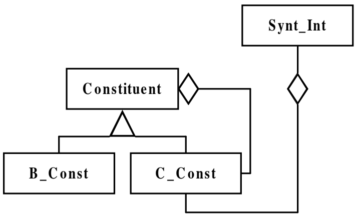


Fig. 1 Syntactic interpretation

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 modeled by the relation between the *c_const* class and the *synt_int* class in the hierarchy.

4.2 The visual representation of annotation

Managing large sentence annotations is cumbersome. Effective graphical representations are needed both for the annotator and the user. Their aim is to ease the navigation in intricate information.

Constituency-based annotation schemes are tree structures. Graphical tree representations aim to ease the user interactions with the tree structures, i.e. the display, retrieval and updating of annotation.

The visual representation defined is a *strip tree* (see Fig. 2) which resembles standard bracketed representations and provides an intuitive and easy to modify hierarchical view of the constituent structure.

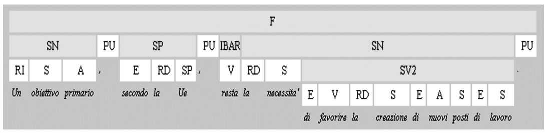


Fig. 2 Strip tree

Functional annotation is visually represented in terms of graphs, where functional relations are drawn as arcs linking the head and the dependent. The insertion/deletion of elliptical material is another essential feature of this tool module.

Finally, lexico-semantic annotation, which proceeds per lemma, does not pose specific representation requirements, while browsing at this level needs the parallel use of the IWN tool.

4.3 *GesTALt* architecture

The *GesTALt* annotation workbench is the resultant system, constituted by a pool of cooperative subsystems. The system manages the linguistic database sketched in section 4.1 and allows its output in the XML standard.

The system is a suite composed by specific applications: *SinTAS* for constituent 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 obtained in *ValTAS* that need all the capabilities spread in the subsystems. The technologies adopted for the development (object-oriented design), in conjunction with an ad-hoc architectural design, allows an easy reuse of the functionalities developed for the subsystems in the global (i.e. *ValTAS*) system.

The overall *GesTALt* architecture is shown in Fig. 3 (overleaf), where components are represented as boxes, and interactions as arrows. The creating/translating flow of the object-oriented database (*GestTALt–OODB*) is shared by the subsystems. Information is extracted from and injected in 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.

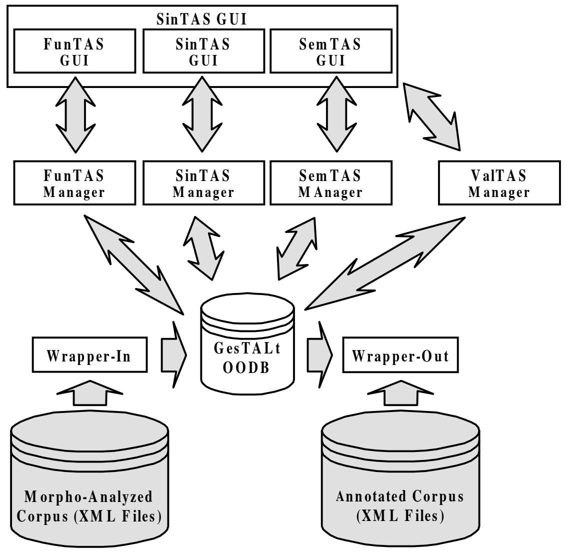


Fig. 3 GesTALt architecture

Each subsystem, but *ValTAS* that include all, is composed by specialized 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*, respectively). Furthermore, the different ways of interaction 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*).

5 Treebank Evaluation

The information stored in ISST, in particular in the financial corpus, will be used to improve an automatic Italian-English translation system, PeTra Word 2.0 , developed by Synthema and already on the market.

PeTra is based on the Logical Grammars ("Slot Grammars") formalism (McCord 1980, 1989) and is composed of three main components: the Italian language analyser (morphologic analyser, monolingual dictionary and syntactic parser), the transfer component (bilingual dictionary and structural transfer rules) and the English morphologic generator. We expect to improve: dictionaries, Italian grammar and transfer rules.

5.1 Changes to the dictionary content

Adding the missing entries: PeTra's dictionary coverage will be enlarged through addition of missing specialised entries and through improvement of already contained entries. Associated translations will be added to the bilingual dictionary.

Inserting new multi-word expressions: the multi-word expressions annotated in ISST will be revised and added to the dictionary either in terms of single entries or of particular constructions associated with component words, considering the system constrains.

Improving lexico-semantic hierarchy: by using lexico-semantic annotation, the semantic-hierarchical dictionary structure will be revised: the semantic attributes are especially used for the lexical transfer disambiguation.

5.2 Analysis Rules

The current grammar has a good coverage (i.e. 88% on unrestricted texts), but it is likely that many structures in the ISST corpus will be analysed incompletely or incorrectly: the corpus is a specialised one and it may contain constructions which are not used in standard Italian. ISST will be examined to check the grammar coverage: accessing ISST on the basis of functional relations, which correspond to the slots, will allow to study the features and the Constituents, in order to determine the possible structures and encode the proper rules.

The translation tests will also allow to determine the sentences which are not recognised by the current grammar: the rules will be modified by retrieving the “similar” structures contained into ISST. The access to ISST will be made through the sentence being examined in order to obtain the two syntactic annotations, study them to determine the uncovered structure and other possible annotations of the same type inside the corpus, and finally analyse them to decide whether and how to apply possible changes.

5.3 Transfer Rules

By analysing all of the new elements included into the analysis rules and revising the translation tests, the set of rules which forms the syntactic transfer can be improved.

5.4 Results Evaluation

The result validation will be made by comparing the translations of texts in the ISST financial corpus. These translations will be obtained before and after the system tuning. The evaluation will verify the improvement obtained.

The software, which will be a support product for the evaluator, will allow to interactively access to the source text and the related translations, and assign a score based on fixed criteria. The evaluation system will also automatically evaluate the amount of unknown words and not closed trees.

References

Abeillé A., Clément L., Kinyon A (2000) *Building a treebank for French*, in Proceedings of LREC-2000, 31/5-2/6 2000, Athens, pp. 87-94.

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

Brants T., Skut W., Uszkoreit H. (1999) *Syntactic annotation of a German newspaper corpus*, in Proceedings of the Treebanks workshop, Journée(s) ATALA sur les corpus annotés pour la syntaxe, 18-19 juin 1999, Université Paris 7, Paris.

Bémová A., J. Hajic, B. Hladká, J. Panenová (1999) *Syntactic tagging of the The Prague dependency Treebank*, in *Proceedings of the Treebanks workshop, Journée(s) ATALA sur les corpus annotés pour la syntaxe*, 18-19 juin 1999, Université Paris 7, place Jussieu, Paris.

Calzolari N. and Corazzari O. (forthcoming) *Senseval/Romanseval: the framework for Italian*. Computers and the Humanities, Dordrecht.

Carroll J., E. Briscoe, A. Sanfilippo (1998) *Parser Evaluation: a Survey and a New Proposal*, in LREC-1998 Proceedings, Granada, Spain, 28-30 May, pp. 447-454.

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

Corazzari O., M. Monachini, 1995, *ELSNET: Italian Corpus Sample*, ILC-CNR, 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*, 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-2000 Proceedings, Athens, June 2000.

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

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

Van Halteren H. (1997) *Excursions into syntactic databases*, Amsterdam, Rodopi.

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

Lenci A., S. Montemagni, V. Pirrelli, C. Soria (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, June 22nd .

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

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

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

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

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

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

Sandoval M., Lopez Ruesga A., Sanchez León S. and F. (1999) *Spanish Tree Bank: Specifications*, Version 4, Manuscript.

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