THE SYNTACTIC-SEMANTIC *TREEBANK* OF ITALIAN
An Overview

SIMONETTA MONTEMAGNI, FRANCESCO BARSOTTI,
MARCO BATTISTA, NICOLETTA CALZOLARI, ORNELLA CORAZZARI,
ALESSANDRO LENCI, VITO PIRRELLI, ANTONIO ZAMPOLLI,
FRANCESCA FANCIULLI, MARIA MASSETANI, REMO RAFFAELLI,
ROBERTO BASILI, MARIA TERESA PAZIENZA, DARIO SARACINO,
FABIO ZANZOTTO, NADIA MANA, FABIO PIANESI,
RODOLFO DELMONTE

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 - *linguistically interpreted corpora, multi-layered linguistic annotation, syntactic annotation, semantic annotation, annotation software, evaluation*

1. 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 fill this gap by developing a multi-level annotated corpus, the Italian Syntactic-Semantic Treebank (henceforth referred to as ISST). A first evaluation of ISST was carried out in the framework of a machine translation application. Specifically developed software, including an intelligent interface, supported both annotation and evaluation activities.

ISST - which represents one of the main actions of an Italian national project, SI-TAL[1] - is developed by a consortium of companies and research institutions in Italy which are active with different expertise in the computational linguistics field.

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 syntax classes at Nijmegen University (Van Halteren, 1997).

2. 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 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

[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).

organisation (e.g. titles, subtitles, summary, body of article, paragraphs).

The multi-level structure of ISST shows two main novelties with respect to other treebanks: 1) 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); 2) 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.

3. ISST INPUT

3.1. *Corpus composition*

The ISST corpus consists of 305,547 word tokens reflecting contemporary language use. It includes two different sections: 1) a "balanced" corpus, testifying general language usage, for a total of 215,606 tokens; 2) a specialised corpus, amounting to 89,941 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).

3.2. *Morpho-syntactic annotation*

Syntactic and lexico-semantic annotation takes as input the morpho-syntactically annotated text. Morpho-syntactic annotation was previously carried out at the Institute of Computational

Linguistics (ILC-CNR, Pisa, Italy) in the framework of the European projects PAROLE (Goggi *et al.*, 1997) and ELSNET (Corazzari and Monachini, 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, as in the case of multi-word expressions.

Multi-word expressions marked at this level are made up of contiguous sequences of words and include: 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 constituted by two prepositions, the first of which also includes the definiteness mark, followed by an adverb); expressions whose grammatical properties do not directly follow from their constituting 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 4.3.1).

4. ISST ANNOTATION SCHEMATA

4.1. *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 both in real applications and for 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.

4.2. *Syntactic annotation*

Most treebanks, currently available or under construction for different languages, adopt a unique syntactic representation layer, following either a constituency-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 two syntactic "views", though complementary, 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 on the developed resource can be obtained, for example, by projecting functional information onto the constituent structure (see section 4.2.3).

This "double-track" approach to syntactic representation is particularly suited to deal with a language like Italian, in particular 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 approx. 70% of the possible cases (Bates, 1976).

These two features combined together would have made a pure constituency-based representation of Italian unrestricted texts quite difficult, due to the massive use of empty elements (either traces or pro-subjects) at the cost of lower transparency of the annotation. ISST overcomes these problems 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 at the level of constituent structure, but only at the functional level.

4.2.1. *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, mainly due to the distributed organisation of syntactic annotation.

Annotation at this level consists in the identification of phrase boundaries with labelling of constituent types. Since 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 è stato risolto IBAR] [COMPT [SP per [SN decreto SN] SP] COMPT] F]`

In the constituent structure in (1), 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 due to its frequent discontinuity, up to the point of becoming controversial whether this notion 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 elliptic element to its governing 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 for example 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 più 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_là_del [SN confine SN] SP] COMPIN] F] F2] SN] COMPT] F]`

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
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

Table 1. ISST constituent types

Constituency annotation in ISST uses an inventory of 22 constituent types, reported in table 1. Following theoretical assumptions derived from the Lexical Functional Grammar theory, syntactic constituents are divided up into functional constituents and substantial constituents. 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). This three way classification is reported in the third column of table 1. Note that 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.

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. Then, the output of the shallow parser is manually revised and corrected.

4.2.2. *Functional annotation*

Functional annotation in ISST is word-based, i.e. it is carried out 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, 2000, 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.

We used FAME (Lenci *et al.*, 1999b, 2000) as the starting point for the development of ISST functional annotation scheme. FAME originates 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), revision which was first done for better complying with the basic requirements of parsing evaluation (in the framework of the LE4-8340 ELSE project, see Lenci *et al.*, 1999a), and then for making the scheme suitable for annotation of unrestricted Italian texts. With SPARKLE, FAME shares two main features: a) it is functional, i.e. the basic units are functional relations holding between lexical heads; b) functional relations are hierarchically organised to make provision for underspecified representations of highly ambiguous functional analyses. The main novelty of FAME stands in its modular architecture: annotation is articulated over different information layers, each factoring out different but possibly interrelated linguistic facets of syntactic annotation. This set of features combined together 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. 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 will be discussed below.

Dependency relations are hierarchically structured to make provision for underspecified representations of highly ambiguous functional analyses. The typology of dependency relations, hierarchically organized, is given in figure 1.

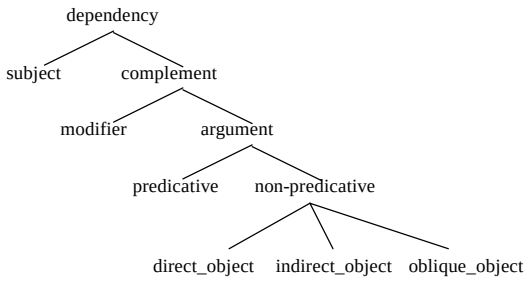

Figure 1. Hierarchical organization of ISST dependency relations

A dependency relation is an asymmetric binary relation between full words, respectively 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 intertranslatability of FAME into other syntactic representation formats see Lenci *et al.*, 1999, 2000).

Annotation at the functional level is carried out manually.

4.2.3. *Relationship between the two syntactic annotation levels*

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' whose constituent structure and functional annotation are respectively reported in (7) and (8) below:

(7) `[F [SN Giovanni SN] [IBAR sembra IBAR] [SV2 arrivare [SAVV domani SAVV] SV2] F]`

(8) `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, where each constituent category is marked, whenever possible, with a functional tag.

(9)
```
[F [SN-sogg Giovanni SN-sogg] [IBAR sembra
IBAR] [SV2-arg arrivare [SAVV-mod domani SAVV-
mod] SV2-arg] F]
```

This is one of the many possible combined views which can be obtained on the ISST syntactically annotated corpus.

4.3. *Lexico-semantic annotation*

4.3.1. *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.*, 2000) 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.

4.3.2. *Annotation criteria*

Each annotation unit is tagged with the relevant sense according to IWN sense distinctions. In order to meet requirement c) in section 4.1 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;
- 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, e.g. title of an opera (e.g. *Il barbiere di Siviglia*), of a newspaper (e.g. *La Nazione*).

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 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.

4.3.3. *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 in figure 2: 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.

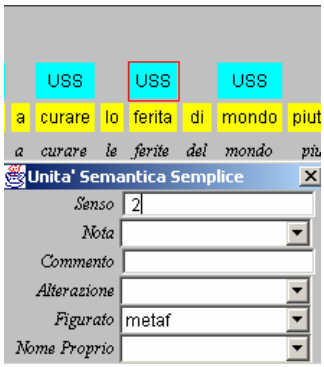

Figure 2. Example of an annotated US

Annotation of semantically complex units (USC) is exemplified in figure 3 for the multi-word expression *essere alle corde* 'to be hard-pressed'. The dark 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).

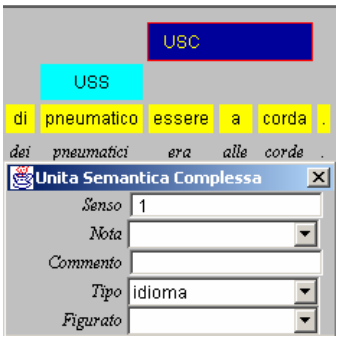

Figure 3. Example of an annotated USC

Finally, an example is given in figure 4 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*).

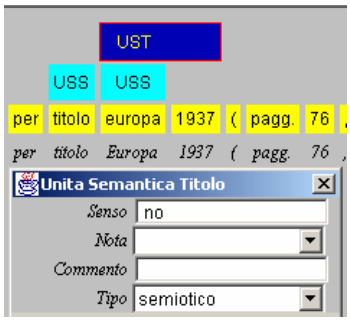

Figure 4. Example of an annotated UST

Obviously, sense information does not apply to UST.

4.3.4. *The added value of corpus annotation*

Sense assignments combined together with the additional lexico-semantic information conveyed through the semantic tags listed in section 4.3.2 make the ISST annotated corpus more than a mere list of instantiations of the senses attested in the reference lexical resource. In this way, the annotated corpus becomes a repository of interesting lexico-semantic information, especially for what concerns lexico-semantic facts which are excluded - either programmatically or just by chance - from the reference lexical resource.

Let us consider the case of word usages which are not lexicalised and, as such, are not recorded in the reference lexical resource. For instance, consider the following contexts, where the target word - marked in bold - is used metaphorically:

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

In both cases, the metaphoric use of target words is specified through a specific tag (FIGURATO=metaf). As to sense assignment, *arma* in (10) is assigned the appropriate figurative sense (IWN sense 2); by contrast, *appuntamenti* in (11), representing an instance of non lexicalized metaphor, is linked to the literal sense. Through the interaction between sense and feature information lexicalised and non lexicalised figurative usages 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 can also be resorted 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:

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

In both (12) and (13) *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 (12) is assigned the 'group of people' value and in (13) the 'place' one.

It may also be the case that corpus annotation identifies multi-word expressions that are not recognized 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. In the examples which follow, identified USC's correspond to the words marked in bold:

(14) ***tagliare le ali*** *a qn* 'to clip somebody's wings'
tagliare le ultime ali *a un paese* 'to clip the last wings to a country'
(15) ***fare affidamento*** *su qn* 'to rely on somebody'
non si deve ***fare troppo affidamento*** *sulla politica monetaria*
'we do not have to rely too much on the financial policy'

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 presents itself as 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 side, this type of corpus annotation can help to enrich or simply tune the reference lexical resource through addition of missing entries (or simply variants) and senses.

5. THE MULTI-LEVEL LINGUISTIC ANNOTATION TOOL

The labour intensive annotation task requires for 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 so 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.

5.1. *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 modeled respectively via the class of objects. To give the reader the flavor of the object modeling 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 5.

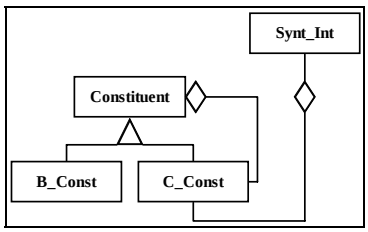

Figure 5. Syntactic interpretation

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.

5.2. *The visual representation of annotations*

Managing the annotation of large sentences is cumbersome. Effective graphical representations are needed both for the annotator and the user to ease the navigation in a complex information network. In what follows, the visual representation of syntactic and lexico-semantic annotations adopted within GesTALt is described.

5.2.1. *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 the user interactions with the tree structures, i.e. the retrieval and the modification of contained information.

Syntactic annotation of unrestricted texts requires to deal with large sentences (of 20 words or more). Sentence length (that determines the number of leaves) combined with the need of 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 6), which is similar to a bracketed representation but gives a hierarchical view of the structured information.

The annotation activity at this level requires to follow the evolution of the tree structures. From this perspective, partial annotations have to be represented, and the transition from a partial annotation to another has to be supported. In figure 6, a partial analysis for the Italian sentence *Un obiettivo primario, secondo la U.E., resta la necessità di favorire la creazione di nuovi posti di lavoro* 'a basic aim, according to E.U., remains the necessity of pushing for the creation of new jobs' is given:

Figure 6. Strip tree (a)

The graphical interface provides easy ways to define constituents via the control of their spanning in the sentence. From our perspective, a partial annotation covers all words of the target sentence.

The interaction with the structure is intuitive. The constituent spans can be clearly defined, and taken decisions (i.e. the defined constituents) produce a clear modification of the *strip tree*. For instance, the definition of a new constituent in the partial annotation in figure 6, i.e. the definition of the SN (noun phrase) *la necessità di favorire la creazione di nuovi posti di lavoro*, requires the identification of the span via a simple interaction. The consequent modification of the tree is depicted in figure 7:

Figure 7. Strip tree (b)

the related chunks of text have been embedded one level down, while no destructive modification was performed.

5.2.2. *Functional annotation*

Dependency-based functional annotations are graphically described 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 4.2.2), another important functionality of this annotation module is constituted by insertion/deletion of nodes corresponding to elliptical material.

An example of functional annotation (sentence (2) above), as it appears from the graphical interface, is given in figure 8:

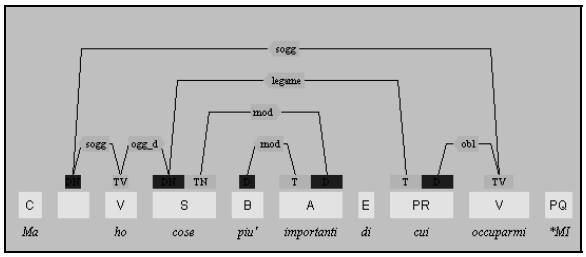

Figure 8. Planar graph

Functional relations are represented as directed labelled arches 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 void boxes: see the null subject in figure 8. The example above also shows the adopted representation of 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 on morphological words (see section 3).

5.2.3. *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 9.

Figure 9. The visual representation of lexico-semantic annotation

The figure shows a semantically tagged sentence where different types of semantic units can be observed. 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 to title units, they can correspond either to individual words or to word sequences; in any case, annotation is both at the level of the composing words and of the title unit. Note that, for each identified semantic unit, annotation is represented in terms of a feature structure specifying the typology of information described in section 4.3 above.

5.3. *GesTALt architecture*

The GesTALt annotation workbench is the resulting system, constituted by a pool of cooperative subsystems. The system manages the linguistic database sketched in section 5.1 and produces its output in standard XML.

The system is a suite composed by 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 obtained in ValTAS that needs 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 figure 10:

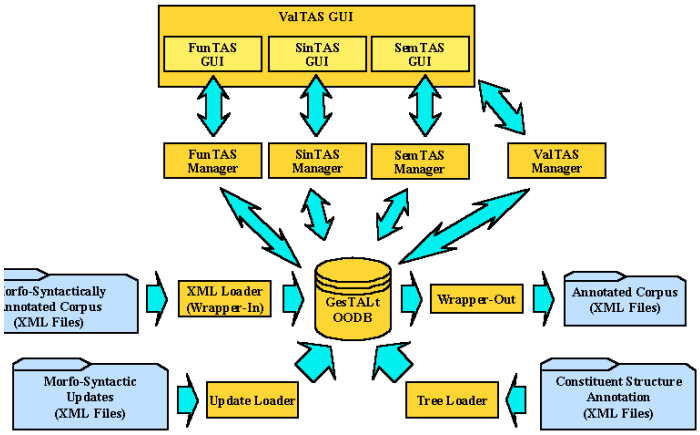

Figure 10. GesTALt architecture

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 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, but ValTAS that includes 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).

6. TREEBANK EVALUATION

The information stored in ISST, in particular 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.

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. Improvements were mainly concerned with mono- and bi-lingual dictionaries, the Italian grammar and the transfer rules.

6.1. *Changes to the dictionary content*

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

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

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

6.2. *Analysis rules*

Before the tuning of the system with ISST, the grammar had already 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 the grammar coverage: by 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 encoded in the rules.

Translation tests also allowed identification of the sentences not recognised by the grammar before tuning: the rules were then modified on the basis of evidence emerging from the analysis of "similar" structures occurring in ISST. Access to ISST was made starting from the sentence being examined in order to retrieve the two syntactic annotations (the functional and the constituency-based ones), study them to identify the structure not covered by the grammar, and finally decide whether and how to apply possible changes to the rules.

6.3. *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 was improved.

6.4. *Evaluation results*

The adopted evaluation methodology can be summarised as follows:

a) identification of a thematically homogeneous subcorpus;
b) translation of the subcorpus before and after tuning;
c) classification of translated sentences into the following typology of cases: correct translation; inaccurate translation (requiring minor revisions); wrong translation; sentence which could not be translated;
d) analysis and comparison of the results obtained before and after the system tuning.

It came out that correctness of results increased of 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 of 45%, and the number of inaccurate translations of 40%. As a consequence, the number of wrong translations decreased of 38%, and of not translated sentences of 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.

7. CONCLUDING REMARKS

The final and 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 three 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 obviously refers to the goals set up within the SI-TAL project, since resources like ISST require a continuous work of refinement and extension. In fact, treebanks and NLP resources in general for their natural vocation should be regarded as open enterprises. Two possible extensions might be envisaged, both along the horizontal dimension and along the vertical one. As to the former, ISST coverage can obviously be extended by adding new annotated texts, also 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).

REFERENCES

ABEILLÉ A., CLÉMENT L., KINYON A., *Building a treebank for French*, in *Proceedings of LREC-2000*, 31/5-2/6 2000, Athens, 87-94.

ALONGE A., CALZOLARI N., VOSSEN P., BLOKSMA L., CASTELLON I., MARTÌ T., PETERS W., *The Linguistic Design of the EuroWordNet Database*, «Computers and the Humanities», Special Issue on EuroWordNet, XXXII (1998), 2-3, 91-115.

BATES E., *Language and Context: Studies in the Acquisition of Pragmatics*, New York, Academic, 1976.

BÉMOVÁ A., HAJIC J., HLADKÁ B., PANENOVÁ J., *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.

BRANTS T., SKUT W., USZKOREIT H., *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, Place Jussieu, Paris.

CALZOLARI N., CORAZZARI O., *Senseval/Romanseval: the framework for Italian*, «Computers and the Humanities», XXXIV (2000), 1-2, 61-78.

CARROLL J., BRISCOE T., CALZOLARI N., FEDERICI S., MONTEMAGNI S., PIRRELLI V., GREFENSTETTE G., SANFILIPPO A., CARROLL G., ROOTH M., *Specification of Phrasal Parsing*, SPARKLE Deliverable 1.1, 1996, also available at the SPARKLE home page at http://www.ilc.cnr.it/.

CARROLL J., BRISCOE E., SANFILIPPO A., *Parser Evaluation: a Survey and a New Proposal*, in *Proceedings of LREC-1998*, Granada, Spain, 28-30 May 1998, 447-454.

CORAZZARI O., CALZOLARI N., ZAMPOLLI A., *An Experiment of Lexical-Semantic Tagging of an Italian Corpus*, in *Proceedings of LREC-2000*, 31/5-2/6 2000, Athens, 691-697.

CORAZZARI O., MONACHINI M., *ELSNET: Italian Corpus Sample*, ILC-CNR, Pisa, 1995.

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

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

DELMONTE R., *Shallow Parsing And Functional Structure In Italian Corpora*, in *Proceedings of LREC-2000*, 31/5-2/6 2000, Athens, 113-119.

GOGGI S., BIAGINI L., PICCHI E., BINDI R., ROSSI S., MARINELLI R., *Italian Corpus Documentation*, LE-PAROLE WP2.11, ILC, Pisa, 1997.

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

KARLSSON F., VOUTILAINEN A., HEIKKILA J., ANTTILA A., (eds.), *Constraint Grammar, a language-independent system for parsing unconstrained text*, Berlin and New York, Mouton de Gruyter, 1995.

LENCI A., MONTEMAGNI S., PIRRELLI V., SORIA C., NETTER K., RAJMAN M., *Corpora for Evaluation*, ELSE (LE4-8340) Deliverable D5, 1999a.

LENCI A., MONTEMAGNI S., PIRRELLI V., SORIA C., *FAME: a Functional Annotation Meta-scheme for Multimodal and Multi-lingual Parsing Evaluation*, in *Proceedings of the ACL99 Workshop on Computer-Mediated Language Assessment and Evaluation in NLP*, University of Maryland, June 22nd 1999, 1999b.

LENCI A., MONTEMAGNI S., PIRRELLI V., SORIA C., *Where opposites meet. A Syntactic Meta-scheme for Corpus Annotation and Parsing Evaluation*, in *Proceedings of LREC-2000*, 31/5-2/6 2000, Athens, 625-632.

LIN D., *A dependency.based method for evaluating broad-coverage parsers*, «Natural Language Engineering», IV (1998), 2, 97-114.

MARCUS M., MARCINKIEWICZ M.A., SANTORINI B., *Building a Large Annotated Corpus of English: The Penn Treebank*, «Computational Linguistics», XIX (1993), 2, 313-330.

MCCORD M.C., *Slot Grammars*, «Computational Linguistics», VI (1980), 31-43.

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

MONACHINI M., CALZOLARI N., *Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora. A Common Proposal and Application to European Language*, EAGLES Recommendations, Pisa, ILC, 1996.

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

SAMPSON G., *English for the Computer*, Oxford, Clarendon Press, 1995.

SAMPSON G., *A proposal for improving the measurement of parse accuracy*, «International Journal of Corpus Linguistics», V (2000), 1, 53-68.

SANDOVAL M., LOPEZ RUESGA A., SANCHEZ LEÓN S. and F., *Spanish Tree Bank: Specifications*, Version 4, Manuscript, 1999.
SINCLAIR J., *The Empty Lexicon*, «International Journal of Corpus Linguistics», 1 (1996), 99-119.
SI-TAL, *Specifiche Tecniche di SI-TAL. Manuale Operativo*, ILC-CNR/CPR, Pisa, 2000.
VAN HALTEREN H., *Excursions into syntactic databases*, Amsterdam, Rodopi, 1997.