

Comparing linguistic information in treebank annotations

Cristina Bosco, Vincenzo Lombardo

Dipartimento di Informatica, Università di Torino
Corso Svizzera 185 - 10149 Torino
{bosco, vincenzo}@di.unito.it

Abstract

The paper investigates the issue of portability of methods and results over treebanks in different languages and annotation formats. In particular, it addresses the problem of converting an Italian treebank, the Turin University Treebank (TUT), developed in dependency format, into the Penn Treebank format, in order to possibly exploit the tools and methods already developed and compare the adequacy of information encoding in the two formats. We describe the procedures for converting the two annotation formats and we present an experiment that evaluates some linguistic knowledge extracted from the two formats, namely sub-categorization frames.

1. Introduction

By providing a very large set of manually labelled parsed sentences, the Penn Treebank has played an invaluable role in enabling the development of state-of-the-art NLP systems. Nevertheless, even if this treebank has allowed in the last few years for a very precise comparison among techniques (e.g. parsing), the strong focalization of training testing and tuning of NLP systems on Penn, has left open several questions on system porting. For instance, a wide variety of experimental evidence supports the idea that results obtained on the Penn Treebank are not reproducible on treebanks of languages other than English, e.g. for German (Dubey and Keller, 2003), Czech (Collins et al., 1999), Italian (Corazza et al., 2004), Chinese (Levy and Manning, 2003), regardless of the increasing availability of annotated materials in Penn-like formats too.

The Turin University Treebank (henceforth TUT) is an ongoing project for building an Italian treebank (the current annotated corpus is downloadable at <http://www.di.unito.it/~tutreeb/>). Beyond developing a resource for Italian (currently limited to corpora of very limited size), the project aims at investigating the causes of the portability of results obtained on the English Penn Treebank on other annotation formats and languages. The project implements, in fact, a multiple format annotation that allows for a faceted comparison among several approaches.

The TUT dependency-based annotation is centered on a notion of grammatical relation, with an augmented (morpho-syntactic-semantic) structure that represents the predicate-argument structures of a sentence (Bosco and Lombardo, 2003), (Bosco, 2004). The TUT annotation aims at capturing the richness of the syntax-semantics interface, which is a crucial layer of representation for several NLP tasks, such as Information Extraction, Machine Translation and Question Answering (Palmer et al., 2005).

Two further formats (Constituency-TUT and Augmented-Penn) have been devised to translate the TUT dependency format into the Penn constituency format, through a cascade of automatic converters. Constituency-TUT is a translation of TUT in a Xbar-like format that annotates the relational information of TUT on phrases. Augmented-Penn is a more flattened, Penn-like constituency-based representation that still annotates the TUT grammatical relations

where possible on phrase labels.

By including both dependency (i.e. TUT) and constituency-based annotations (e.g. Constituency-TUT, Augmented-Penn and standard Penn), and also functionally richer (e.g. TUT) and poorer schemes (e.g. Penn), the project allows for pinpointing representation problems and testing the adequacy of information encoding also with reference to different languages (i.e. English and Italian).

The next section is an overview on the treebank different formats. In the third section, these formats are described with some more details together with the relative annotation/conversion processes. The fourth section is finally devoted to the validation of the project with reference to a task of linguistic knowledge extraction.

2. Overview on TUT project

2.1. Data

TUT corpus currently includes 1,800 annotated sentences, which correspond to 52,755 tokens (including words and parts of compound words). The largest portion of the corpus consists of texts from newspapers (50%) and from the code of the Italian civil law (40%); the rest (10%) is a collection of miscellaneous texts from novels and academic prose.

2.2. TUT annotations at glance

The four different annotation formats implemented by the TUT project are exemplified in figure 1, which shows the syntactic trees for the sentence "Il governo di Berisha appare in difficoltà" (The government of Berisha appears in trouble). Starting from the first, which is TUT (figure 1a), these formats show different layers of variation/similarity with respect to the last one, which is standard Penn (figure 1d), in terms of both richness of functional-syntactic information (i.e. amount and specificity of grammatical relations) and type of linguistic framework (i.e. constituency versus dependency, or minimal versus maximal projection). The first format, (a), is standard TUT format, a dependency-based annotation. Each node represents a word, while dependency relations label edges of the tree, in order to make explicit the relation between pairs of words. Non-terminal nodes are not allowed in such a kind of tree. The second format, (b), is the Constituency-TUT format, an

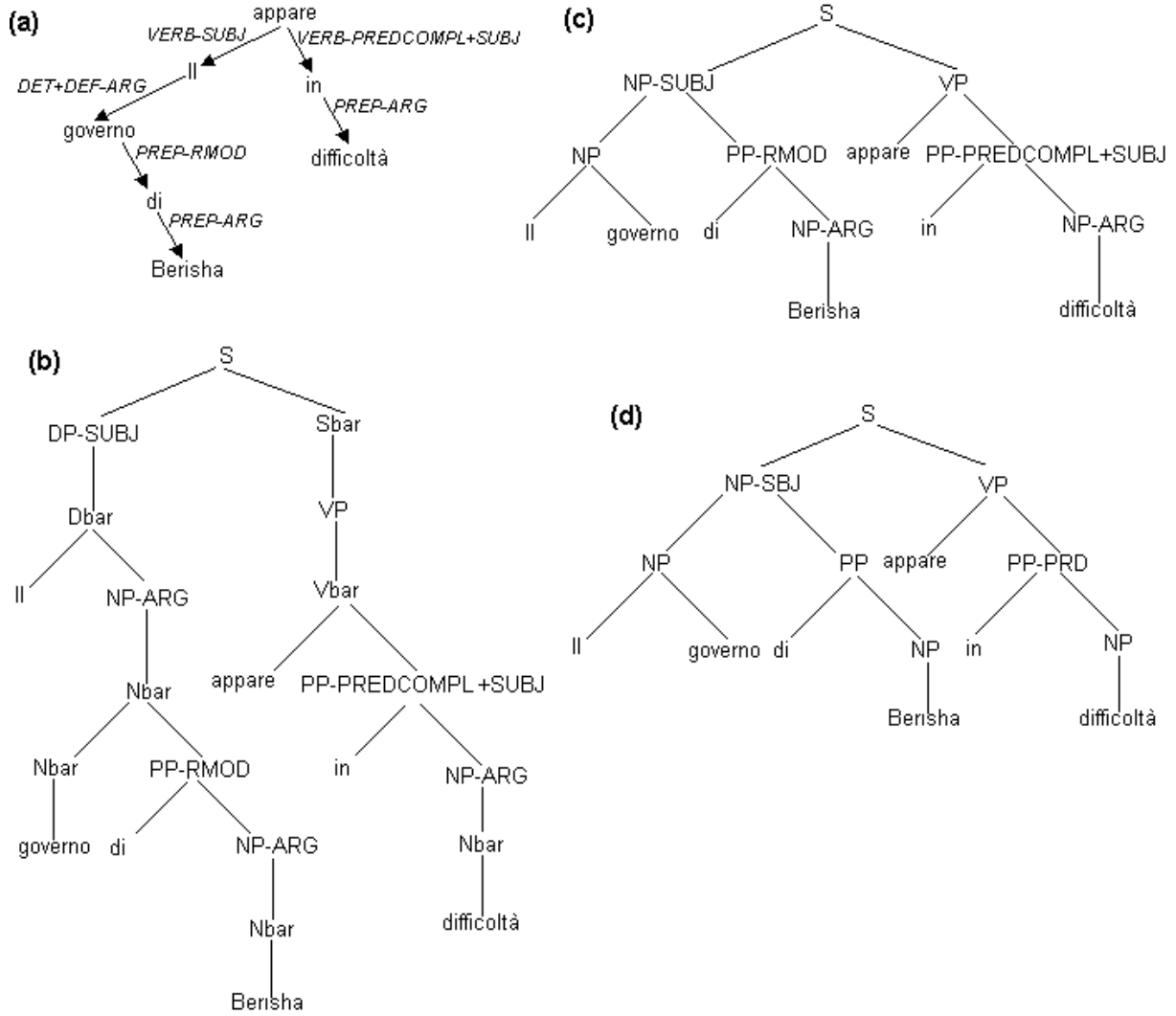


Figure 1: TUT (a), Constituency-TUT (b), Augmented-Penn (c) and Penn-like (d) representations of “Il governo di Berisha appare in difficoltà” (The government of Berisha appears in trouble)

annotation where terminal nodes represent words and non-terminal nodes represent the syntactic sub-units of the sentence, i.e. constituents which are grammatical category projections. Following the major tenets of the Xbar theory (see e.g. (Radford, 1997) and (Santorini, 2000)), Constituency-TUT applies a maximal projection strategy, i.e. all grammatical categories project intermediate and maximal projections (e.g. Verb projects first in Vbar and then in VP; Noun projects first in Nbar and then in NP). Rather than on edges, here, the functional-syntactic relations are annotated on constituents.

The format (c) is Augmented Penn. On the one hand, the Augmented Penn still annotates the dependencies of TUT (when possible on constituents), thus showing a repertoire of functional relations larger than standard Penn. On the other hand, its trees are structurally identical to those of Penn. In fact, like Penn, it applies a minimal projection strategy that produces trees that are flatter than Constituency-TUT by allowing only in some cases for intermediate projections, i.e. a terminal category projects only when the projected constituent includes more than one

word. Moreover, in this format various kinds of structures are standardized according to Penn, e.g. the repertoire of constituents is smaller than in Constituency-TUT and does not include the determiner phrase DP.

The last format, (d), is standard Penn, which includes a few functional relations and implements a minimal projection strategy. In practice, the trees are the same of the Augmented Penn, but relations annotated on constituents are fewer, e.g. OBJ is not marked.

The next section describes the four formats in more details, together with the relative annotation and conversion processes.

3. Annotation and conversion

In order to increase the possibilities of comparison and evaluation of the TUT corpus, we have converted the dependency annotation format into other annotation formats thus yielding a number of parallel treebanks based on the same corpus of sentences. Parallel treebanks may serve as a suitable infrastructure for the comparison of parsers from different linguistic frameworks (Musillo and Sima'an,

2002).

3.1. TUT dependency annotation

TUT dependency-based annotation follows the major tenets of Hudson's Word Grammar (Hudson, 1984). The partial configurationality of Italian, usually assumed in literature and confirmed, e.g., in (Bosco and Lombardo, 2004), is a motivation for the choice of such a dependency-based annotation for this language.

Instead, the richness of the grammatical relation system implemented by TUT is motivated by the aim of satisfying the requirements imposed by NLP tasks, and in particular the representation of the predicate-argument structure (see e.g. (Palmer et al., 2005)) which is crucial for several NLP tasks. TUT annotation distinguishes and encompasses in a single layer three kinds of information involved in grammatical relations, i.e. dependencies, by means of the Augmented Relational Structure (ARS) (Bosco and Lombardo, 2003), (Bosco, 2004). The ARS is a representation of grammatical relations as tripartite structures, complex objects that take into account various interrelated informational domains, called components, i.e. morpho-syntactic, functional-syntactic and semantic-syntactic. The morpho-syntactic component consists in the morphological categories of the words involved in the relation, such as PREP and VERB in figure 1 (a) within the labels for the relations respectively linking the Preposition with its argument ("di" (of) with "Berisha", and "in" (in) "difficoltà" (trouble)), and the Verb with its subject ("appare" (appears) with "Il governo" (The government)) and its predicative complement (with "in difficoltà" (in trouble)). The functional syntactic component distinguishes among a variety of dependency relations, such as SUBJ and ARG in figure 1 (a), e.g. within the labels for the relations respectively linking the Verb with its subject ("appare" with "Il governo") and the Preposition with its argument ("di" with "Berisha"). The syntactic-semantic component discriminates among different kinds of adjuncts and oblique complements, such as TIME and MANNER. Valid tags for the morpho-syntactic component are 40, for the functional-syntactic are 55, and for the semantic-syntactic one they are 88.

Moreover, TUT implements a trace-filler mechanism for representing pro-drop or equi phenomena, but also extractions and long distance dependencies.

The annotation process of TUT is semi-automatic. It includes a Part of Speech tagging (Boella and Lesmo, 1998) followed by a parsing and a manual correction. Finally, after the solution of possible inter-annotator disagreements, the annotated materials are checked first by a tool (that finds errors in the tree structures, such as cycles, crossing edges and unconnected nodes), and then manually corrected. No automatic control based on linguistic knowledge is currently implemented, but an indirect very fine-grained detection of annotation errors is offered by the conversion tools that we describe in the rest of this section.

3.2. From TUT format to Penn format

The conversion of an existing treebank in a popular format has made technique developed for that format ap-

plicable to new corpora, languages, tasks. For instance, the Prague Dependency Treebank (Böhmová et al., 2003) has been converted into a Penn format, allowing for the development of a statistical parser for Czech (Collins et al., 1999); the Penn Treebank has been converted into the Prague format (Zabokrtsky and Kucerova, 2002); the Arabic Penn treebank has been converted into a dependency format (Zabokrtsky and Smrz, 2003).

Nevertheless, even where treebanks are available, NLP techniques developed for English trained and tested on the Penn Treebank, produce worse results than on Penn Treebank and for English, e.g. in parsing Czech (Collins et al., 1999), German (Dubey and Keller, 2003), Italian (Corazza et al., 2004) and Chinese (Levy and Manning, 2003). Therefore, the conversion into Penn format is furthermore crucial for investigating portability of NLP models from English to other languages. It allows for checking robustness of linguistic information encoded in annotations and comparing different schemata, to see whether they can be improved and what is the relation of schemata and traditional linguistic information employed in conversion procedures.

In the TUT project, in order to smooth the conversion process, we have devised other two intermediate formats (see figure 2). In this paper we focus on the first step, i.e. the conversion from TUT to Constituency-TUT.

3.2.1. Constituency-TUT

Constituency-TUT is a TUT-oriented constituency-based annotation with TUT relations annotated on constituents. Each terminal category X corresponds to a node (i.e. word) of a TUT tree, and projects into non-terminal nodes which represent intermediate (Xbar) and maximal (XP) projections of X, according to Xbar theory (Radford 1997). The distinction between complements and adjuncts is structurally marked according to the Xbar theory; therefore complements, usually closer to their head, are daughters of intermediate projections and sisters of terminal categories, while adjuncts are both sisters and daughters of intermediate projections (e.g. [XP (Xbar (Xbar (X)(COMPLEMENT)) (ADJUNCT))]).

The conversion algorithm from TUT to Constituency-TUT is adapted from that in (Xia, 2001) for the conversion of dependency structures, i.e. ordered dependency trees¹, in Penn-like phrase structures, i.e. constituents featuring a minimal projection strategy. The input of our implemented algorithm are the TUT ordered dependency trees where the explicit marking of null elements is allowed, whilst the output are constituency trees applying a maximal projection strategy. The main information that is present in a constituency tree but not in a dependency tree is the type of the

¹A dependency tree is ordered if the dependents of the same head are ordered according to their positions in the sentence and the structure is projective. A very few non-projective structures have been found in TUT. Moreover, since empty nodes are allowed in TUT dependency trees, cases of non-projectivity have been dealt with by adding empty nodes in the sentence representation. Thus, by explicitly adding empty nodes in TUT dependency trees, we can represent also inherently non-projective sentences through projective structures.

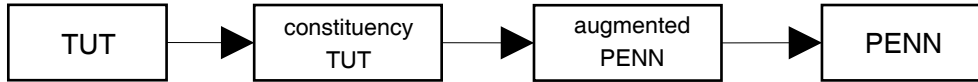


Figure 2: The conversion schema.

multiple word syntactic units (e.g. NP, VP and S). Therefore, the major goal of the algorithm is to recover the types of phrases that each node of the dependency tree projects. In other words, what are the expansions in constituency terms of the grammatical category of terminal nodes which have to be annotated as non-terminal nodes in the constituency trees. Moreover, in a dependency tree several grammatical categories interact and in order to build a corresponding constituency tree we have to know in which way we can represent this interaction in constituency terms, that is how grammatical categories and their projections combine in a constituency structure. This is a language specific information that consists in listing the types of adjuncts and complements that a head can take and their positions with respect to the head itself.

3.2.2. Augmented-Penn

Augmented-Penn is a constituency-based representation that annotates TUT relations, and features, like Penn, minimal projection, i.e. each terminal category projects only when the constituent includes more than one word. It results in a Penn format more functionally annotated.

Two are the main goals of the algorithm converting Constituency-TUT to Augmented-Penn. The first consists in applying the minimal projection strategy, i.e. exploiting a sort of pruning of intermediate projections of grammatical categories. The second consists in standardizing the repertory of phrases according to the Penn format. There is, in fact, a few specific structures which show minor conceptual differences in TUT with respect to Penn which are mirrored in Constituency-TUT too. These differences are smoothed away by the converter from Constituency-TUT to Augmented-Penn by modifying the structures where these differences occur. For instance, in Penn, in Verbal complexes the finite tensed Verb governs the infinite Verbs, e.g. in the verbal structures including auxiliary Verbs, the tensed auxiliary Verb governs the main Verb (and other auxiliary, if any); by contrast, in TUT the un-tensed main Verb is the head of the verbal structure and the auxiliaries depend on the main Verb.

3.2.3. Penn

Penn is the well known format of the Penn Treebank. The syntactic annotation of this treebank is constituency-based, but enriched with the annotation of a small set of grammatical relations and semantic information². It makes easy

the automatic recover of a skeletal notion of predicate argument structure (e.g. subject, but not object are explicitly annotated as grammatical relations) and the structure of discontinuous constructions (e.g. by using a trace-filler notation). In fact, each constituent has zero or more 'function tags' indicating semantic roles and other information related to syntax but not encapsulated in the simple constituents. Each constituent can be tagged with multiple function tags, but never with two tags of the same kind. The conversion from Augmented-Penn to Penn, mainly deletes the relations not in use in the PT. Since the set of TUT relations is richer than that of Penn, the conversion makes the output functionally poorer than the input.

4. Using TUT

Parallel annotations for the same corpus of sentences may serve as a suitable infrastructure for comparisons among different linguistic frameworks. The definition of a conversion process from a format A to another format B is in itself a comparison between these formats. In fact, in order to develop the conversion, we have to describe a virtually complete and correct mapping which translate every analysis in the treebank A into the corresponding analysis in the linguistic framework B.

In this section we present an experiment that refers to the automatic acquisition of lexical knowledge about Verbs, and more precisely sub-categorization frames. Since the annotation of TUT is especially centered on predicate-argument structure and features a detailed representation of verbal complements and adjuncts, thus the experiment focusses on a specific issue on which the TUT formats are meaningfully comparable. Moreover, Verbs are the primary source of relational information in the sentence, and the lexical knowledge about Verbs is critical for a wide range of NLP tasks, such as parsing (Collins, 1999) or machine translation (Surdeanu et al., 2003). In particular, sub-categorization frames encode the syntactic correlate of the semantic predicate-argument structure associated with verbs that relates an action or a state to its participants.

The experiment consists in a comparison among data extracted from TUT and those extracted from a manually constructed commercial Italian dictionary. We started by extracting from the 1,800 sentences of TUT (i.e. 52,755 tokens) each Verb occurring in active form: 3,711 tokens (Verb forms) are extracted and classified in 830 types (Verb lemmas). Table 1 shows the distribution of lemmas versus forms in TUT, with a few Verbs occurring very often (e.g. avere (to have) and essere (to be)) and most Verbs occurring one or two times only.

We extracted for each token the sub-categorization frame

²Penn functional tags are:

- grammatical tags: DTV = dative, LGS = logical subject, PRD = predicate, PUT = locative complement of "put", SBJ = surface subject, TPC = topicalized, VOC = vocative
- form/function tags: ADV = adverbial, NOM = nominal
- semantic role tags: BNF = benefactive, DIR = direction,

EXT = extent, LOC = location, MNR = manner, PRP = purpose and reason, TMP = temporal.

Verb Lemmas	Verb forms
401	1
147	2
62	3
59	4
34	5
14	6
13	7
10	8
14	9
19	10
13	11
4	12
5	14
3	15
5	16
2	18
2	20
5	21
1	24
1	25
2	26
1	27
1	30
1	34
1	38
1	41
1	116
1	123
1	269
1	442

Table 1: Distribution of active Verb lemmas versus forms in TUT corpus.

associated in the corpus, e.g. for an occurrence of the Verb "vedere" (to see) in active transitive form we can find the frame *VEDERE:left(VERB-SUBJ),right(VERB-OBJ)*.

Then we compare this output with the sub-categorization information extracted from the digitalized Italian dictionary DISC (Sabatini and Coletti, 1997). Within DISC we selected only the information concerning the number (or numbers) of arguments that a Verb accepts. DISC includes the sub-categorization frames concerning 9,884 different lemmas of Italian Verbs. Some obsolete form or reflexive Verb lemma occurring in TUT does not occur in DISC too (e.g. "rinunziare" (to renounce) which is an obsolete Verb that occurs in TUT but not in DISC, where occurs "rinunciare"), therefore the comparison referred only to 3,691 tokens rather than 3,711. In table 2 are the results of this comparison between sub-categorization frames extracted from TUT and those in DISC.

The results show that with the relational information (i.e. dependencies) annotated in TUT 94,77% of tokens match with the DISC data. An error analysis makes clear some major issues that require further study. In particular, the difficulty of predict the intransitive use of transitive Verbs and alternation of verbal patterns, in order to distinguish cases

Active Verbs	matching Disc	unmatching Disc
3,691	3,498 (= 94,77%)	193 (= 5,23%)

Table 2: Comparison between Verb sub-categorization frames extracted from TUT and those extracted from an Italian dictionary.

of real object drop Verbs from cases where a transitive Verb is not legitimate at stay without the object. For instance, "giocare" (to play) in Italian like in English can take or not the object; instead, "avvertire" (to warn) in the sentence "La Blue Guide di James Pettifer sull'Albania avverte" (The Blue Guide of James Pettifer concerning Albania warns) does not instantiate a right sub-categorization frame, may be because of a missing context³.

Moreover, since the search performed by the extractor program mainly depends on the annotated dependencies and their labels, e.g. the distinction between complements and adjuncts and among various kinds of oblique complements too, we can predict that similar results on the same task can be achieved on other equally functionally rich formats, such as Constituency-TUT, but not by functionally poorer formats, like the Penn one. It is, in fact, well known in literature that this kinds of extraction from e.g. Penn Treebank has to be driven by heuristics which are limited since tuned on the specific corpus, see e.g. in (Collins, 1999). Only a certain amount of information concerning predicate argument structure can be automatically determined, as the need for the PropBank demonstrates (Palmer et al., 2005). Automatic predicate argument analysis are reported in literature, like those presented in (Palmer et al., 2001) which obtain 83,7%, (Gildea and M., 2002) which obtain 82,8% and (Xue and Kulick, 2003) which obtain 95%. But the results are not directly comparable with ours for several reasons. First, (Palmer et al., 2001) and (Gildea and M., 2002) refer to richer lexical databases featuring a sort of ontology (i.e. VerbNet and FrameNet respectively) which include types and number of elements sub-categorized by Verbs, thus allowing for the development of automatic tagger of predicate argument structure. Second, Italian Verbs are probably associated with fewer senses than English; therefore the mapping between predicate argument structure is probably easier, but this issue requires further investigation.

The possible directions for the future development of our research are many. Three are the major kinds of data that we actually excluded from our research and have to be included in order to extend the completeness and thus comparability of this preliminary research: passive forms of Verbs, Verb nominalizations (such as destruction for destroy), adjuncts participating in predicative structures. In particular, since the major goal is the comparison with Propbank (Palmer et al., 2005) and the Chinese PropBank (Xue and Kulick, 2003), we have to develop the research by referring to more richer lexical resource including ontological organization of Verbs, their complements and adjuncts.

³As usual in treebanks, in TUT there are not intra-sentential annotations.

5. Conclusions

The paper has presented a treebank project for Italian that includes the conversion from the original dependency format to the Penn format, through passages of intermediate formats. Then we have seen an experiment on the extraction of linguistic knowledge that aims at comparing the subcategorization frames extracted from the dependency format and the Penn format respectively, thus showing the better accuracy of richer formats. The side effect of the conversion process has been an increased check of the annotation format.

6. Acknowledgement

We would like to acknowledge the Giunti Multimedia publisher, for making available the DISC subcategorization frames used in this research.

7. References

- G. Boella and L. Lesmo. 1998. Automatic refinement of linguistic rules for tagging. In *Proceedings of the LREC'98*.
- Alena Böhmová, Jan Hajič, Eva Hajičová, and Barbora Hladká. 2003. The Prague Dependency Treebank: A three level annotation scenario. In Anne Abeillé, editor, *Building and using syntactically annotated corpora*. Kluwer, Dordrecht.
- C. Bosco and V. Lombardo. 2003. A relation-schema for treebank annotation. In A. Cappelli and F. Turini, editors, *Advances in Artificial Intelligence, LNCS 2829*.
- C. Bosco and V. Lombardo. 2004. Dependency and relational structure in treebank annotation. In *Proceedings of the Workshop on Recent Advances in Dependency Grammar at COLING'04*.
- C. Bosco. 2004. *A grammatical relation system for treebank annotation*. Ph.D. thesis, University of Torino.
- M.J. Collins, J. Hajic, L. Ramshaw, and C. Tillmann. 1999. A statistical parser of Czech. In *Proceedings of the ACL'99*.
- M.J. Collins. 1999. *Head-driven statistical models for natural language parsing*. Ph.D. thesis, University of Pennsylvania.
- A. Corazza, A. Lavelli, G. Satta, and R. Zanolli. 2004. Analyzing an Italian treebank with state-of-the-art statistical parser. In *Proceedings of TLT-2004*.
- A. Dubey and F. Keller. 2003. Probabilistic parsing for German using sister-head dependencies. In *Proceedings of the ACL'03*.
- D. Gildea and Palmer M. 2002. The necessity of syntactic parsing for predicate argument recognition. In *Proceedings of the ACL '02*.
- Richard Hudson. 1984. *Word Grammar*. Basil Blackwell, Oxford and New York.
- R. Levy and C. Manning. 2003. Is it harder to parse Chinese, or the Chinese treebank? In *Proceedings of the ACL'03*.
- G. Musillo and K. Sima'an. 2002. Towards comparing parsers from different linguistic frameworks. an information theoretic approach. In *Proceedings of Workshop Beyond PARSEVAL - Towards improved evaluation measures for parsing systems at the LREC'02*.
- M. Palmer, Rosenzweig J., and Cotton S. 2001. Automatic predicate argument structure analysis of the Penn Treebank. In *Proceedings of the Human Language Technology (HLT) 2001*.
- M. Palmer, D. Gildea, and P. Kinsbury. 2005. The Proposition Bank: and annotated corpus of semantic roles. *Computational Linguistics*, 31(1).
- A. Radford. 1997. *Syntactic theory and the structure of English. A minimalist approach*. Cambridge University Press, Cambridge.
- Sabatini and Coletti. 1997. Disc compact Dizionario Italiano Sabatini Coletti. CD-ROM edition.
- B. Santorini. 2000. The syntax of natural language: an online introduction using the Trees program. <http://www.ling.upenn.edu/beatrice/syntax-textbook/index.html>.
- M. Surdeanu, S. Harabagiu, J. Williams, and P. Aarseth. 2003. Using predicate-argument structures for information extraction. In *Proceedings of the ACL '03*.
- F. Xia. 2001. *Automatic grammar generation from two different perspectives*. Ph.D. thesis, University of Pennsylvania.
- N. Xue and S. Kulick. 2003. Automatic predicate argument structure analysis of the Penn Chinese Treebank. In *Proceedings of the Machine Translation Summit IX (MTIX) 2003*.
- Z. Zabokrtsky and I. Kucerova. 2002. Transforming Penn Treebank phrase trees into (Praguan) tectogrammatical dependency trees. 78. <http://ckl.mff.cuni.cz/zabokrtsky/ws2tgts/>.
- Z. Zabokrtsky and O. Smrz. 2003. Arabic syntactic trees: from constituency to dependency. In *Proceedings of the EACL'03*.