Comparing treebanks

Marietta Sionti
University of Athens

Abstract

In order to significantly promote the study of Natural Language Processing phenomena towards the understanding of text and spoken language, computational linguists attempt to automatically extract lexical information from very large corpora. These annotated corpora have already been applied to research in a wide range of disciplines from speech recognition to theoretical linguistics. Specifically, textual data contribute to training statistical models for the development of typical theories to create several grammars, as well as for the evaluation and comparison between sufficient analysis models and other applications (Marcus *et al.*1993). The necessity of building these types of resources has led us to the development of such annotated corpora—namely treebanks—for different languages, e.g. English, German, Czech, Italian among others. This paper attempts to compare these treebanks in terms of formalism, language, levels of annotation, annotation method and manipulation of specific phenomena, such as discontinuous constituents, ellipsis and coordination. Although the paper is mainly focused on the abovementioned comparison, our final goal is to conclude with a formalism and solutions, which can be best adopted for the Greek language. The final outcome is to propose a Greek Treebank, which emphasizes the annotation of sentences with relatively free word order and complex noun phrases as arguments.

1 Introduction

The aim of the present work is to study and compare several treebanks, namely functionally and structurally annotated corpora (i.e. Treebank of the University of Pennsylvania, Prague Dependency Treebank, Turin University Treebank), in order to examine their features and propose a Greek Treebank Scheme based on the novelties of the existing treebanks. The Greek Treebank Scheme focuses on the annotation of sentences with relatively free word order as well as on complex noun phrases that are used as arguments in the frame of coordination (Sionti 2008).

As far as the understanding of text and spoken language is concerned, treebanks are significant tools in achieving a both robust and rapid development in the study of NLP phenomena, especially when the automatic extraction of lexical information from various corpora is attempted. Such annotated corpora-treebanks, enriched with a deeper possible analysis, have been offered as accurate NLP research tools, facilitating thus voice and speech recognition while assisting theoretical linguistics as well. In particular, textual data have been already applied in statistical models training (empirical approach) for the development of various formal theories, the appraisal of analytical models' suffiency and in innumerous other applications (Marcus *et al*. 1993).

Generally, treebank schemes include at least two distinctive annotation levels. The first one is called 'part of speech tagging'. Usually, this is a completely automated process during which the lexical–morphological information is marked. The second, 'syntactic tagging', presents some form of depiction of the sentence's syntactic relations. Depending on the treebank, syntactic characteristics i.e. number, gender and time, as well as semantic aspects may appear in tree structures.

Descriptiveness and data-drivenness are general specifications of such annotation schemes (Skut *et al*. 1997).

- Annotation, according to descriptiveness, aims exclusively at describing linguistic phenomena that appear in texts instead of explaning them.
- Annotation extracts conclusions from data (data-drivenness).

The design of a scheme for linguistic annotation acknowledges the need for compromising between theoretical linguistics descriptions and practical goals, which should be served by the annotation scheme. Therefore, whenever a theoretical description of a language is proposed, linguists are moved not only by theoretical assumptions but also by their linguistic intuition in their attempt to describe in detail syntactic structures and relationships. On the contrary, a treebank is a functional annotation model, which is designed for specific technological applications. Therefore, detailed theoretical accuracy has to be partly sacrificed for the sake of robustness of practical applications; choices should primarily aim at theoretical correctness and descriptive adequacy.

2 Treebank of the University of Pennsylvania (UPENN treebank)

One of the first and most innovative treebanks is the one developed at the University of Pennsylvania where most of the input consists of written texts taken from the Wall Street journal and the Brown corpus. In the early years of the project, brackets were only used, in a merely skeletal structure, while in later stages UPENN adopted a richer format that combined the use of brackets with the predicate-argument structure. In the latter phase, an annotated version of the Switchboard corpus—a transcript of telephone conversations—was produced.

2.1 Methodology

The three types of UPENN annotation are: (i) part of speech tagging, (ii) syntactic tagging, (iii) disfluency annotation. These are produced by the same method in two stages, while automatic annotation is followed by manual correction.

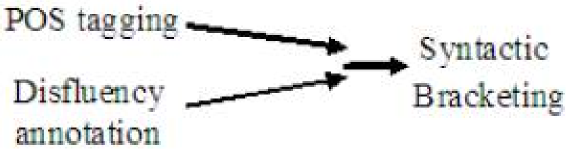


Figure 1. Merging both results for extracting syntactical bracketing

2.1.1 1st Level: Part of Speech tagging (POS)

In the early stages of the UPENN treebank, the initial automated POS form was produced by the PARTS stochastic algorithm, which allowed a 3-5% error rate. Whenever a rule-based stochastic parser was engaged, the error was reduced to 2%. Finally, it was replaced by the Brill syntactic parser.

2.1.2 2nd level: Syntactic Annotation

In the first phase, corpora were annotated with syntactic bracketing with the use of Fidditch, a deterministic algorithm which, due to its properties, is ideal for pre-processing manual annotation in a short time and producing large-scale annotated texts, at the same time.

After the first UPENN Treebank version was released, many users pointed out that it still needed:

- Richer and deeper annotation
- Increased consistency of the introduced corpus
- Less skeletal annotation form
- Expanded context free analysis to cover discontinuous structures.

All the needs could be covered, if a predicate - argument structure was used. So the second UPENN Treebank version came out. The new annotation approach contained the following novel information:

- a simpler annotation engine that provided a clear distinction between arguments and adjuncts.
- a context free mechanism that was easier to find structures with discontinuous constituents.
- a series of empty elements.

```
(S (NP-SBJ (NP It)
 (S * EXP *- 1))
 (VP is
 (NP a pleasure))
 (S-1 (NP-SBJ *)
 (VP to
 (VP teach
 (NP her )))))
 Predicate Argument Structure:
 pleasure (teach (* someone *, her))
```

Figure 2. Syntactic annotation structure

Predicate - argument structure aims to assign an appropriate semantic label to each argument, which would be derived by the predicate. In addition, the above mentioned structure distinguishes arguments from adjuncts. Unfortunately, while such distinction is quite easy in these simple cases, which serve as examples, it is considered as a particularly difficult problem in cases of texts.

2.2 Discontinuous constituents

We can tackle the majority of the trapping and discontinuous constituents phenomena using simple methods, such as co-indexing, in order to index discontinuous structures. In this case, a numeric index (index number) is added to both the label of the original component and the empty element. Sequentially, this number will determine the position of the element in the predicate-argument structure.

The UPENN treebank uses a variety of empty elements to highlight the dependences of discontinuous constituents. Literature refers to such components naming them pseudoattached and distinguishes four different types.

1. Label * ICH *
2. Label * PPA *
3. Label * RNR *
4. Label * exp *

```
(S      (NP-SBJ Chris)
        (VP KNEW
                (SBAR * ICH * -1)
         (NP-TMP yesterday)
                (SBAR-1 that
 (S     (NP-SBJ Terry)
        (VP would
        (VP catch
                (NP the ball )))))))
```

Figure 3. Example of empty elements

2.3 Ellipsis- Empty categories

The main innovation brought out by the UPENN treebank - compared to the Lancaster Project - is the annotation of the "empty" elements in a wide variety of situations. In the following example, the empty elements are co-indexed in the referred lexical entity. UPENN uses symbols for the gaps:

* T * indicates the WH movement and topicalization

* The remaining empty categories are accomplished with the insertion of an integer to the non-terminal category (e.g. NP-10, VP-25) which is used as the component's identification. Sequentially, the empty element is indexed with the same integer. For indexing WH-questions, the UPENN uses SBARG and SQ for auxiliary inverted structures. Only in cases of WH-movement, it uses WH-(NP - PP) tags, which always leave a co anaphora trace.

```
(SBARQ (WHNP-1 what)
(SQ is
(NP-SBJ Tim
(VP eating
(NP * T * - 1)))
```

The predicate argument relation is extracted from the above structure, if the empty element is replaced with the referred word: Predicate Argument Structure: eat (Tim, what).

3 Prague Dependency Treebank (PDT)

The Prague Dependency Treebank has three levels:

- The lower level, which is called morphological.
- The medium level is called analytical and annotates the surface structure with the use of Dependency Grammar.
- The upper level is called tectogrammatical or the level of linguistic concept.

The texts used for PDT, selected by the Czech corpus, contain:

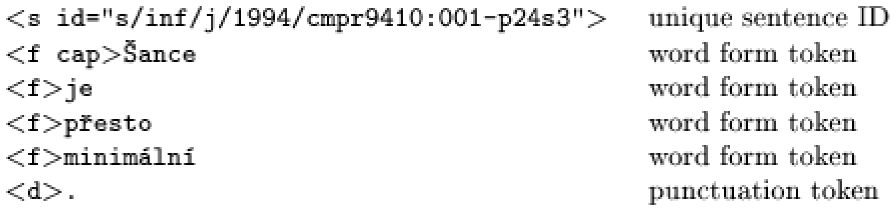
1. Articles of general interest (60% politics, sports, culture, etc.)
2. Financial news and analysis (20%)
3. Popular Science Magazines (20%).

PDT is a two-phase long-term program; the first, deals with the first two annotation levels and the second, deals with grammatical annotation.

3.1 Morphological level

In case of ambiguity, the morphological analysis of a single word produces one or more lemmas. The combination of the value is called morphological tag (MTag). The list of possible MTags with the corresponding entries represents the morphological analysis of imported forms of words. In a given context, there is only one couple (MTag, lemma).

Sequentially, a morphological annotated corpus can train a syntactic parser based on a probabilistic model, which can then be used to annotate a large number of new texts.



```
<s id="s/inf/j/1994/cmpr9410:001-p24s3">   unique sentence ID
<f cap>Šance                               word form token
<f>je                                      word form token
<f>přesto                                  word form token
<f>minimální                               word form token
<d>.                                       punctuation token
```

Figure 4. Morphological tagging (PDT)

3.2 Analytical level

The analytical level is the second level of the general schema. The Dependency Grammar represents the surface syntactic relations of a sentence. The tree structure is based on a relationship of parent or dependent node.

The basic principles of the analytical level are:

- every word and every punctuation mark are represented by only one node,
- no node is inserted (with the exception of a special "technical" node used for the tree's root)
- each node of the resulting analytical tree consists of three parts:
  (i) the original form of the word

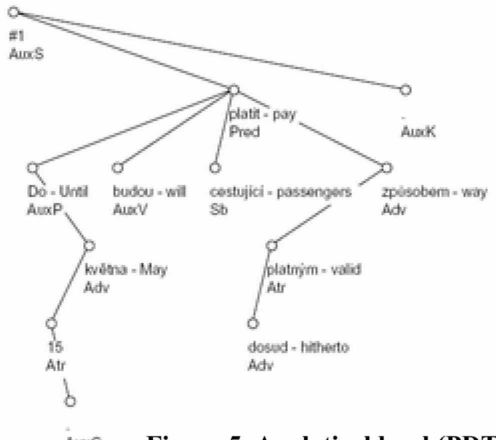


Figure 5. Analytical level (PDT)

(ii) the morphological tag and its lemma (derived unchanged from the morphological level)

(iii) the syntactic label Stags

Do 15 kveta budou cestujici platit dosud platnym zpusobem.

Until May 15th, passengers will be paying using the current scheme.

Table 1. Values of function afun-the most representative one of the analytical level) (Bohmova *et al.*, 2000)

afun	Description
Pred	Predicate if it depends on the added root node (main predicate)
Sb	Subject
Obj	Object
Adv	Adverbial (without a detailed type distinction)
Atv	Complement; technically depends on its non-verbal governor
AtvV	Complement; if only one governor is present (the verb)
Atr	Attribute
Pnom	Nominal predicate's nominal part
AuxV	Auxiliary Verb \to be"
Coord	Coordination node
Apos	Apposition node
AuxT	Reflexive particle, lexically bound to its verb
AuxR	Reflexive particle, which is neither Obj nor AuxT (passive)
AuxP	Preposition, or a part of compound preposition
AuxC	Conjunction (subordinate)
AuxO	Referring particle or emotional particle
AuxZ	Rhematizer or other mode acting to stress another constituent
AuxX	Comma (but not the main coordinating comma)
AuxG	Other graphical symbols, not classified as AuxK
AuxY	Other words, such as particles, without specific (syntactic) function, parts of lexical idioms, etc.
AuxS	The (artificially created) root of the tree (#)
AuxK	Punctuation at the end of sentence, or direct speech, or citation clause
ExD	Ellipsis handling
AtrAtr	A node (analytical function: an attribute) which could depend also on its governor's governor
ObjAtr	There must be no semantic or situational difference between the two cases

3.3 Tectogrammatical level

The third level (tectogrammatical) describes the meaning of a sentence, presents the structure of the sentence and uses Dependency Grammar.The representation of the tectogrammatical level is analogous to the analytical tree. Each label consists essentially of two sets of properties:

- a word's lexical value (the properties of lemma: t-lemma)
- the variables corresponding to the syntactic functions

In PDT, the problem of textual ellipsis is solved in the analytical and tectogrammatical levels. At the analytical level, it is shown with the ExD label at the remaining node but in this case, there are different solutions for different elliptical phenomena such as the ellipsis of prepositional phrases (which take the AuxP label) or subordinated sentences (AuxC). Each label consists essentially of two sets of properties:

- the word's lexical value (the properties of lemma: t-lemma, the so-called morphosyntactic grammateme), that reflects the concept of morphological categories.
- the variables that correspond to the editorial functions

In addition to the above mentioned variables, PDT is employing subtle differentiation of syntactic relations by means of so-called agreement grammateme (for this work, there were used 12 grammatemes).

4 Turin University Treebank (TUT)

Experience showed that, patterns based on Dependency Grammar (DG) are more appropriate for languages with relatively free word order because their functional role does not depend directly on the word's position in the sentence (Skut *et al.* 1997).

The Italian language is characterized by laxity in the position, especially of the verb. So, a Dependency Grammar based schema, which combined the predicate-argument structure, was developed.

TUT displays descriptive richness and flexibility to the sentence

- Dependency relations render prolificacy to the scheme. These relations are already inherited from the annotational phase, while their presence is imposed by the need for a close to semantics representation.
- Flexibility results from the hierarchical organization of dependency relations, from general to specific. The upper levels of the taxonomy present the main types of relationships, e.g. predicate. The lower levels point out more specific relations e.g. prepag (prepositional predicate).

4.1 Levels of annotation

The levels of annotation are two; morphological and syntactical. The first level, similar to other treebanks, indicates tree nodes with morphological tags. This is a semi-automatic process where initially a morphological parser analyzer is attributing tags to the words and then human annotators correct the outcome of the software. Respectively, the syntactical level is more complex than in other tree banks. In this level, grammatical relations are incorporated, in addition to the syntactical-functional ones.

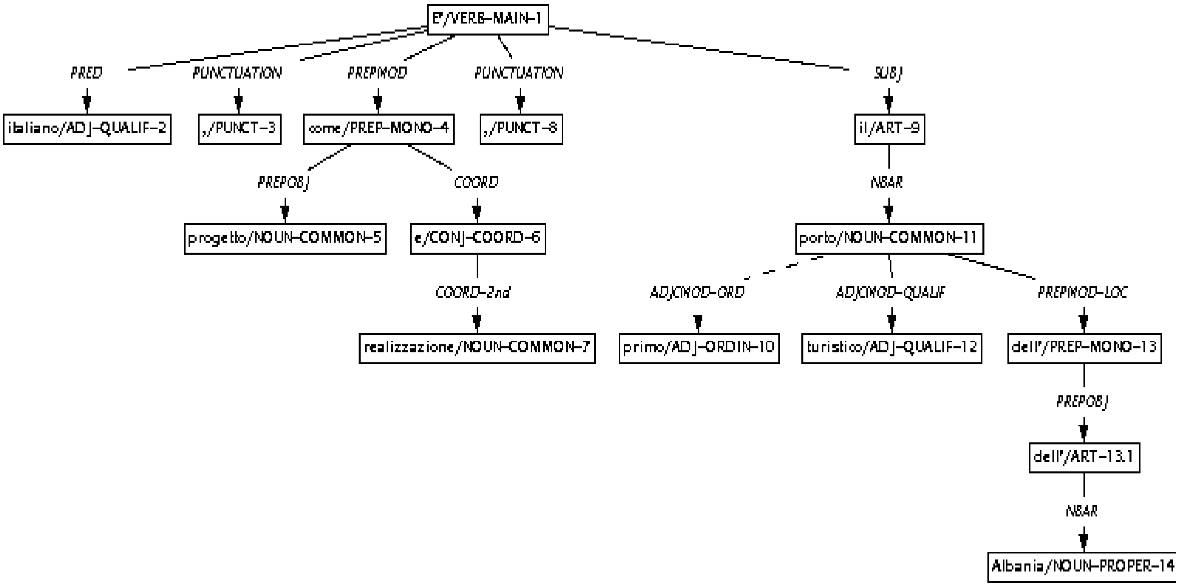


Figure 9. Syntactical scheme of the sentence *E Italiano, come, progetto e realizzazione, il primo porto turistico dell 'Albania*

For accuracy reasons, since discontinuous elements are quite rare in the Italian language, there is no risk of trace overgeneration in the syntactic tree. The second advantage resulting from the presence of traces in the Italian formalism refers to the allowance of a clear separation between continuous and discontinuous dependencies.

Table 2. Summary table of Treebanks (Sionti 2008)

	UPENN	**PRAGUE**	**TURIN**
Formalism	Phrase structure and Predicate-argument structure	Dependency Grammar	Dependency Grammar
Language	English Binding word order	Czech Free word order	Italian Free word order
Levels	Part of speech tagging I) Syntactic tagging (with parenthesis) II) Predicate argument structure annotation	Morphological level Analytical level Tectogrammatical level	Morphological level Syntactic level enriched with grammatical relations
Annotation method	Semi-automated	Semi-automated	Semi-automated
Discontinuous constituents	Traces	Traces	Traces
Ellipsis	Traces	Traces	Traces
	Charniak, 1996; Marcus *et al.*, 1993, 1994 ftp://ftp.cis.upenn.edu/pub/treebank/doc/manual	Bohmova, 1999; Hajic,1998 http://ufal.ms.mff.cuni.cz/pdt/pdt.ht	Bosco, 2000; 2001 www.di.unito.it

5 Comparison

In general, a treebank contains at least two distinct layers of annotation. The first layer, annotating the parts of speech (POS tagging), is usually a fully automated process during which words appear with morphological information. The second layer, called syntactic annotation, consists of a specific representation of the sentence's syntactic relations. Some treebanks assigned grammatical features (number, gender, time, etc.) and other semantic aspects in the tree structures (Marcus *et al.* 1993, Moreno & Lopez 1999).

Except from descriptiveness, an annotation schema requires data-driveness, which implies that conclusions are drawn from data (Skut *et al.* 1997). Descriptiveness is associated with the commentary object, which should be described-rather than explained- by linguistic phenomena in the corpus. Conclusions drawn from the data provide the necessary tools for all types of grammatical patterns. Furthermore, a treebank must achieve a consistent treatment of similar grammatical structures. Such consistency is reflected in the uniform manner used by the treebank in describing the same phenomenon (Marcus *et al.* 1994). The consistent and clear distinction between different levels of feedback will prevent redundancy and duplication of information. The selective treatment of the tree banks material is allowed. Table 2 presents the main features of the Penn Treebank for English (Marcus *et al.* 1994), the Prague Treebank for Czech (Bohmova *et al.* 1999) and spots the major differences among them.

6 The Greek Treebank scheme

Since word order in Greek demonstrates a particular laxity,[1] Dependency Grammar (Kakouriotis 1988) appears to be the preferred formalism. Priority would be given in phrase structure only if we were interested in the allocation of the words within the sentence, in the event that this constituted the main factor of syntactic characterization, as in English. In Greek, however, elements' functional roles do not depend directly on the word's placement within the sentence (except for tight structure cases in the NP).

Furthermore, the Greek language possesses a large number of discontinuous components, thus releasing us from the need to mention the location of each component. So the syntactic annotation of linguistic data becomes feasible, without requiring the existence of empty elements for the reservation of place or other theoretical changes in the data. Nevertheless,

[1] The Greek language demonstrates great laxity as far the word order is concerned, due to the components' rich inflectional system. The recognition of syntactic relations between textual elements is feasible while their placement is not bound at fixed places, from which they will finally get their functional roles. A simple declarative sentence with a verb, as well as nominal subjects and objects, can be rewritten in six likely combinations (Georgiafentis, 2004):

1. O	Petros	ide	to	ergo	SVO
The.NOM	Peter.NOM	saw.3SG	the.ACC	play.ACC	
2. Ide	o	Petros	to	ergo	VSO
saw.3SG	the.NOM	Peter.NOM	the.ACC	play.ACC	
3. Ide	to	ergo	o	Petros	VOS
saw.3SG	the.ACC	play.ACC	the.NOM	Peter.NOM	
4. TO	ERGO	ide	o	Petros	OVS
THE.ACC	PLAY.ACC	saw.3SG	the.NOM	Peter.NOM	
5. TO	ERGO	o	Petros	ide	OSV
THE.ACC	PLAY.ACC	the.NOM	Peter.NOM	saw.3SG	
6. O	Petros	TO	ERGO	ide	SOV
The.NOM	Peter.NOM	THE.ACC	PLAY.ACC	saw.3SG	

The above-mentioned possible structures neither have the same rate of appearance nor are pragmatologically equivalent (Laskaratou 1994). Nevertheless, only the first three (SVO, VSO and VOS) are neutral and are characterized as basic. The other three reallocations (OVS, OSV and SOV) function emphatically, emphasizing the informative value of any term (Cleris 1996). According to Laskaratou's research, the Greek language's statistically prevailing structure was traditionally the SVO.

discontinuous components deserve respect because of their impact on the semantic and discourse annotation level.

The Greek Scheme could borrow principles from the Prague Dependency Treebank (PDT), having adopted elements from the Turin University Treebank (TUT). As far as the scheme's general structure is concerned, all three above-mentioned treebanks exploit three annotational levels. In the Greek Treebank we could use the following levels accordingly: morphological, syntactical and structural–functional.

The existence of the first two 'inferior' levels (morphological and syntactical) is a given due to fundamental morphological and syntactical linguistic information. At the 'superior' (structural–functional) level, we will attempt to incorporate not only the whole semantic knowledge (the so-called θ-roles, elements for the achievement of an agreement and overall grammatical relations) but also any individual syntactical adjustments made in order to overcome particular problems (e.g. in ellipsis, the phenomenon will be tackled in a unified manner at the syntactic level while any differences, which depend on the type of ellipsis, will be attributed to the structural–functional level).

Such a form of information could be incorporated in the syntactic level but this would be particularly complicated, from the point of view of both software and human annotator. The latter is expected to extend to all phenomena assisted by complicated instructions.

Should, however, these levels be separated in syntactic and structural–functional, then the human annotator will be able to handle them faster and more effectively (with fewer errors); initially, the first level according to its requirements and then the following one in line with its guidelines and specializations. Finally, the human annotator will be enabled to suggest easily a credible solution for the same or similar problems from an already inferior and more accessible level, preventing any type of interference from other particularities and thus offering transparency and accessibility to the annotation (Sionti 2008).

References

Bohmova, A., Hajic J., Hajicova E. and B.Hladka. 1999. *The Prague Dependency Treebank.*

Bosco, C. 2000. An Annotation Schema for an Italian Treebank. In *Proceedings of the ESSLLI-2000 Student Session,* 22–33

Bosco, C. 2001. Grammatical Relations System in Treebank Annotation. In *Proceedings of Student Research Workshop of Joint ACL/EACL Meeting*, ed. by E. Mitsakaki, C. Monz and A. Ribeiro, 1–6

Charniak E. 1996. *Treebank Grammars*. Brown University.

Cleris, Ch. and G. Babiniotis. 1996. *Modern Greek Grammar: the noun* [in Greek]. Ellinika Grammata.

Georgiafentis M. 2004. *Focus and word variation in Greek.* PhD Thesis, University of Reading.

Hajic, J. 1998. Building a Syntactically Annotated Corpus, Prague Dependency Treebank. In *Issues of Valency and Meaning. Studies in Honor of Jarmila* Panevova, ed. by E. Hajicova, 106–132. Charles University Press.

Kakouriotis, A. 1988. *Dependency Syntax: a comparative approach to the study of English and Modern Greek.* Thessaloniki.

Laskaratou, C. 1994. Performance Principles in word order variation. In *Parousia Journal Monograph Series No 29.* Athens.

Marcus, M., G. Kim, M. A. Marcinkiewicz and R. MacIntyre, A. Bies, M. Ferguson, K. Katz and B. Schasberger. 1994. The Penn Treebank: Annotating Predicate Argument Structure. In *ARPA Human Language Technology Workshop.* Princeton.

Marcus, M.P., G.Kim and M.A.Marcinkiewicz. 1994. The Penn Treebank: Annotating Predicate Argument Structure. In *Proceedings of the Human Language Technology Workshop*: Morgan- Kaufmann.

Marcus, M. P., B. Santorini, and M.A. Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. In *Computational Linguistics* 19.313–330.

Sionti, M. 2008. A theoretical approach towards and practical solutions for a Greek Treebank Scheme. In *New perspectives in Greek linguistics,* ed. by N. Lavidas, E. Nouchoutidou and M. Sionti. Cambridge Scholars Publishing.

Skut, W., T. Brants, B. Krenn and H.Uszkoreit. 1997. An Annotation Scheme for Free Word Order Languages.In *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP).* Washington, D.C.

Prague Dependency Treebank. http://ufal.ms.mff.cuni.cz/pdt/pdt.ht, retrieved 09/20/2008

Turin University Treebank www.di.unito.it, retrieved 09/20/2008

UPENN ftp://ftp.cis.upenn.edu/pub/treebank/doc/manual/

www.cis.upenn.edu/~treebank, retrieved 09/20/2008