Parallel multi-theory annotation of syntactic structure

Jerid Francom and Mans Hulden

The University of Arizona
Tucson, AZ 85721
{jeridf,mhulden}@email.arizona.edu

Abstract

We present an approach to creating a treebank of sentences using multiple notations or linguistic theories simultaneously. We illustrate the method by annotating sentences from the Penn Treebank II in three different theories in parallel: the original PTB notation, a Functional Dependency Grammar notation, and a Government and Binding style notation. Sentences annotated with all of these theories are represented in XML as a directed acyclic graph where nodes and edges may carry extra information depending on the theory encoded.

1. Introduction

Following the establishment of various treebanks for an assortment of languages, the past few years have been marked by strong interest in the development of parallel treebanks—that is, treebanks that simultaneously annotate sentences in two or more languages. For machine translation purposes in particular, this effort seems to prove quite valuable. In contrast to such multi-language annotation work, far less attention has been given to *multi-theory* annotations: annotating treebanks in parallel following different linguistic theories.

The current work highlights limitations in current treebank formulations that potentially restrict the accessibility and overall usefulness of treebanks and suggests that such limitations can be reduced by incorporating syntactic annotation for a set of sentences in multiple theories. Furthermore, a multi-theory treebank creates a host of novel opportunities that cannot be addressed under current approaches to corpus annotation. Yet the practical problems surrounding such multi-theory treebanks are quite different in nature from those seen in multi-language treebanks. In what follows, we shall present the arguments supporting the creation of a multi-theory treebank, an elaborated review of the properties to consider in annotating sentences in a variety of theories simultaneously, provide a set of examples in various theoretical frameworks, and discuss the difficulties one faces in the development of such a multi-purpose treebank.

2. Treebanks and linguistic theory

Treebanks such as the Penn Treebank (Marcus et al., 1994) have become a rich source of data applicable to research in theoretical, corpus and computational linguistics. The syntactic annotation supplied in these sources allows the researcher the ability to extract sub-sentential information from a range of sentence types for a wide variety of applications.

However, the current set of treebanks is not without shortcomings. First, very few treebanks in wide use today overlap with respect to the collection of sentences that have been annotated.[1] The data used in a given treebank may not be optimal for all types of research due to the nature of the source.[2] This, in turn, hinders the potential of any given treebank to serve as a general-purpose resource. Also, many treebanks either assume a theory-neutral approach to annotation or adopt some very specific theoretical framework. In some cases, the lack of detailed theory-specific information has been overcome by adding more refined annotations to existing generic treebanks (Honnibal and Curran, 2005; Hockenmaier and Steedman, 2007). Also, particular theoretical questions may not be easily pursued with existing treebanks due to the specific needs of the researcher, despite the attractiveness of the data on other grounds (i.e. text-source, language, size, etc.) Taken together, the overall usefulness of treebanks may not be fully realized under the current formulation.

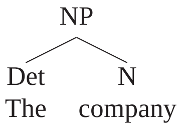
Our proposal aims to fill this apparent gap and create novel research opportunities by elaborating an encoding scheme for a ‘parallel multi-theory’ treebank—that is, a treebank which simultaneously annotates sentences in two or more theoretical frameworks. With the popularity of *multi-language* treebanks and the growth of more reliable parsing methods that require less manual intervention than before, the possibility of annotating a single collection of sentences using different theories is becoming feasible with much less effort than in the past.

A multi-theoretical approach to treebanks shows promise in overcoming those aspects of current treebanks that restrict their overall convenience as a tool for research. This approach creates an alternative to the theory neutral vs. theory specific issue of annotation while providing the ability to extrapolate information over the same data set regardless of the researcher’s theoretical flavor of interest.[3] Furthermore, given explicit correspondences between theoretical machinery, a multi-theory collection also provides a host of other potential applications that benefit from the opportunity to ask questions about the data in one theory *relative* to

[1] The larger richly annotated treebanks available for English such as the Penn treebank (Bies et al., 1995) and LinGO Redwoods (Oepen et al., 2004) all annotate different sets of sentences.

[2] Treebanks vary on a variety of properties that may render an particular dataset less-than-optimal, including the text source: newspaper, web, conversation, etc.; language, relative size, etc.

[3] Overall comparability is enhanced by maintaining the same set of sentences across the different theories, whereas treebanks that work with different sentences are not essentially as directly comparable.



```
<NP>
  <Det><word="The"></Det>
  <N><word="company"></N>
</NP>
```

Figure 1: *A potential simple XML approach that allows only phrase structure modeling.*

another theory, or another version of the same theory, that are not feasible without the elaboration of this ty pe of data set.

Advances in parallel treebank construction demonstrate the possibility of parallel annotation of a single collection of data (Samuelsson and Volk, 2005). However, the encoding of a collection of the sort proposed here poses practical questions about how to maximize its usefulness that are somewhat distinct from those found in multi-language treebanks. At their most basic level both a multi-language and multi-theory treebank must encode sub-sentential units and the relative correspondences between each of these in a parallel dataset. Yet in addition to this level of description and the challenges posed therein, a multi-theory treebank must also create an exchange format that is capable of encoding more detailed properties of a linguistic theory in general and the nuances of particular theories while at the same time providing the substrate for further theoretical augmentation. A more detailed description of the properties to model and the encoding schema employed in this appr oach follow.

3. Properties to model

In encoding our exchange format in XML for simultaneous annotation of sentences in multiple theories we have tried to exploit the following common ground between most theories: trees, dependency structures, collections of features and values, etc., can all be considered elements of a directed acyclic graph (DAGs). A standard phrase-structure tree, as seen in figure 1, *could*, of course, for the most part, be encoded directly into XML by embedding one structure type inside another. However, by doing so, one would lose the ability to handle crossing constituents, dependency grammars, and the like. As noted by Mengel and Lezius (2000), this can be remedied simply by using the XML format to specify a number of nodes and words, which may be linked to each other—thus allowing one to represent crossing dependencies.

In addition to simply regarding a theoretical framework as a collection of nodes and edges, one would also like to retain the hierarchical structure imposed by formats such as XML in order to take advantage of the myriad of tools available for handling XML-encoded data. Also, in order to be as flexible as possible, a multi-theory annotation should share as many generic theoretical devices as possible, while allowing for differences only when absolutely required.

As an exploratory framework, we have chosen three different theories to encode simultaneously: the standard Penn Treebank II format (Bies et al., 1995) (from which our sentences were also taken), functional dependency grammar (FDG) (Tapanainen and Järvinen, 1997), and the generative Government and Binding theory (where we assume Haegeman (1994) to be the standard GB reference).

```
<s id="s1">
  <words>
    <word id="w1_0" orth="The"/>
    <word id="w1_1" orth="company"/>
    <word id="w1_2" orth="can"/>
    <word id="w1_3" orth="go"/>
    <word id="w1_4" orth="about"/>
    <word id="w1_5" orth="its"/>
    <word id="w1_6" orth="business"/>
  </words>
  <theories>
    ...
  </theories>
</s>
```

Figure 2: *The basic XML encoding for individual words.*

Together these representations span a variety of theoretical machinery that needs to be captured in an annotation. While the PTB structure is almost purely phrase-structure based, GB theory requires the modeling of a number of non-phrase-structure elements: movement and traces, co-indexation, theta-roles (through which subcategorization information is expressed), and feature-value combinations, such as those relating to case, person, number, etc. The functional dependency grammar again requires that it be possible that a set of words depend on each other in various ways; dependency grammars generally rely on classifying dependencies between words into different *types* as opposed to phrase structure grammars where there is only a single type of edge between nodes (representing a parent-child node relation).

Despite this plurality, we have made the following simplifying assumptions to bring different models as close together as possible:

- Phrase structure trees and dependency structures are DAGs with nodes possibly having edges to the orthographic representation of a string of words
- Nodes and edges may carry a set of properties that are theory-specific (although, for FDG, the only edge-related property used is that of "label").

3.1. Example annotations

In our model, every sentence shares the same set of words—i.e. the orthographic representation, which can be seen as a preamble to the more theory-specific annotations that follow. For instance, the sentence from the PTB *The company can go about its business* (see figure 6 for the PTB representation), would entail an encoding as in figure 2.

The identifiers of the words simply serve as possible targets for nodes in the more theory-specific section that follows the `<words>` section. This orthographic section is followed by a specific section for each theory included in the encoding. The following `<theory> ...`

</theory> section is largely identical in format for all theories as regards nodes and edges. However, each <node> section contains a theory-specific section defining the properties of a particular node as illustrated in the following examples.

In figure 3, we see a GB tree of the above example sentence with the XML-representation of node *I* illustrating a) the various "properties" the node carries: tense, person, number, co-indexation, labeling, and b) an edge from the node to the orthographic index, and c) how the phrase-structure tree is encoded by directed edges.[4]

Similarly, figure 4 illustrates the corresponding FDG-annotated graph, where the node associated with the first word 'The' is seen in XML.[5] Despite the superficially different theories of GB and FDG, the only difference in this FDG encoding and the GB one is is that a) an edge in FDG can carry a "label" property, and b) the entities in the <properties> section are only of two kinds: functions, and parts-of-speech (pos).

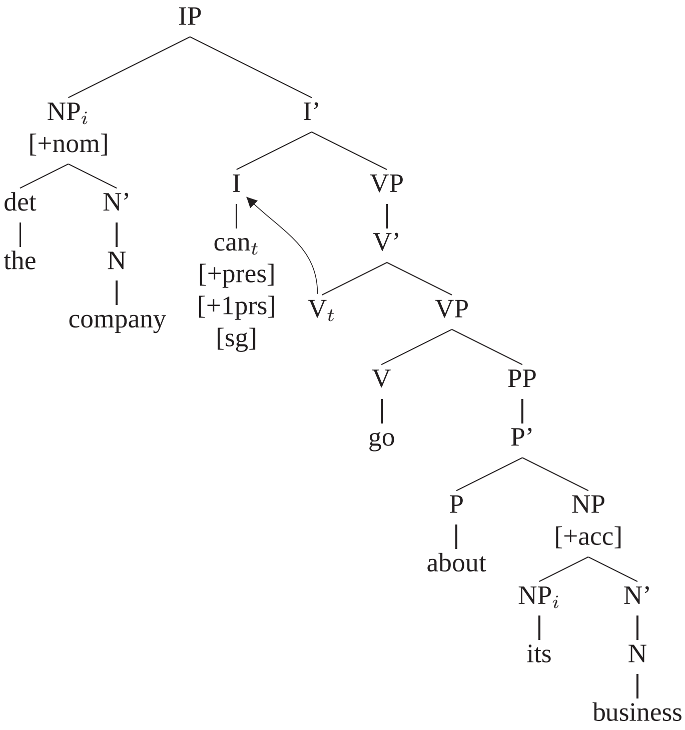
For the sake of completeness, figure 6 contains the original Penn treebank sentence encoded according to the same scheme.

3.1.1. A note on the definition of 'word'

The method used here, where words in the orthographic sentence carry a one-to-many relationship with node elements, is required for cases where a theory splits up the orthographic word into several elements. The inverse relationship, a many-to-one relationship from words to nodes is required when a theory collapses several words into one node. Both scenarios are somewhat rare in English (but frequent elsewhere) as most major theoretical approaches have a rough agreement on the "status" of the word unit. However, a simple example of the first type in English is found in negation contractions, where a *'t* element is considered a separate word in many theories, but is contracted with an auxiliary in the orthographic representation. In this case, the orthographic word may be pointed to by two different nodes in a tree.

The second type of relationship could be the result of an analysis of phrasal verbs (as the *go about* example used earlier)—where a phrasal verb consisting of two separate orthographic words is associated with only one node. Obviously, both scenarios are encodable without allowing for edges from words to nodes, but simply by declaring that a node may consist of several edges that connect to words, and that many nodes can connect to just one word.

Whether or not this approach provides a sufficiently fine-grained annotation obviously depends on the theory and the language: for a highly agglutinating language where a long word may be pointed to by a large number of nodes, one may want to add a morpheme-layer between the level of node and word, where a node would point to a morpheme, and a morpheme be a part of a word. For our purposes (dealing with English primarily) we have refrained from this additional layer for the sake of simplicity.



```
<theories>
  <theory_gb>
    <node id="n1_5" ref="w1_2">
      <properties>
        <property="label" value="I"/>
        <property="tense" value="pres"/>
        <property="person" value="1"/>
        <property="number" value="sg"/>
        <property="index" value="1"/>
      </properties>
    </node>
   ...
   </theory_gb>
...
</theories>
```

Figure 3: *The example sentence in GB theory with encoding details shown for the node **I**.*

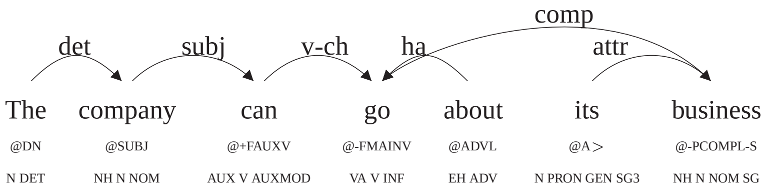
4. Potential applications

As discussed in Section 2., a multi-theory approach to treebanks alleviates some of the limitations associated with 'traditional' treebanks by providing a more general-purpose data source. However, we also anticipate a number of fruitful applications of multi-theory treebank annotations that either enhance existing enterprises or break new ground where investigation was not possible.

In corpus research, it may be the case that no one particular theory is fine-grained enough in some limited area to provide a method for searching for some phenomenon of interest. This issue then lies in the particular perceived lim-

[4] Note that the trace and corresponding movement of the V-node is captured by the index number on that node, which will also be present on the lower V$_t$ node. The information about the presence of a trace and its movement is thus recoverable from the indices, and need not be listed as a separate property.

[5] Sentences from the Penn treebank were analyzed for FDG with Connexor's Machinese "tagger".



```
<theories>
  <theory_fdg>
    <node id="n1_0" ref="w1_0">
      <edge id="e1_1" ref="n1_1"
       label="det"/>
      <properties>
        <property="function"
         value="@DN"/>
        <property="pos" value="N␣DET"/>
      </properties>
    </node>
    ...
  </theory_fdg>
  ...
</theories>
```

Figure 4: *The example sentence in FDG with encoding details shown for the node associated with the word **The**.*

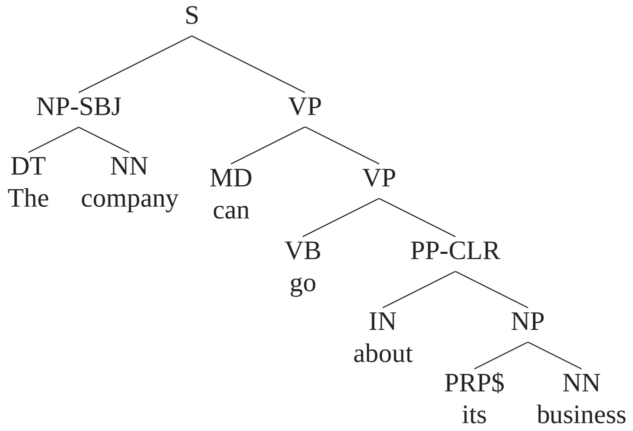


Figure 5: *The example sentence from the original PTB.*

```
<theories>
  <theory_ptb2>
    <node id="n1_0" ref="w1_0">
      <edge id="e1_1" ref="n1_1">
      <properties>
        <property="label" value="S">
      </properties>
    </node>
    ...
  </theory_ptb2>
  ...
</theories>
```

Figure 6: *The example sentence in PTB II with encoding details shown for the node **S**.*

itations of the theory (not the treebank) in question. In this case, allowing for combined queries over several different syntactic annotations may be profitable in that the combined properties may in fact produce nuances and search possibilities that otherwise would not be easily formulated. Also, for grammar induction purposes, we foresee that offering a selection of annotations to choose from and test over may be of interest. Recent works on grammar induction on the PTB have chosen an interesting approach to discard the original functional tags (such as NP-SBJ and PP-CLR in figure 6) because of their perceived "negative utility" in the learning task. Instead of the original tags, the bare syntactic tags have been enriched through what is called "parent annotation," first introduced in Johnson (1998). Somewhat surprisingly, this kind of automatic "functional tagging" to replace the original PTB tags has turned out to perform very well, as shown in Klein and Manning (2003), who attain nearly state-of-the-art parser performance without lexicalization through a few systematic modifications to the original annotations. Along these lines, we also see automatic enrichments on existing treebanks to be a possible target for paral lel annotation, where one could annotate a collection of sentences with minor variants of the same notation simultaneously.

Maybe most the notable novel application that a multi-theory treebank provides for is found within theoretical linguistics. One can construct tools for theoretical research which would allow a user to query a treebank about how some element in theory X is represented in theory Y. Using such a tool, researchers would be in a unique position to evaluate particular theoretical implications of a given property and evaluate the coverage of a grammar. Indirectly, a collection of this sort encourages researchers to formalize proposed theories more rigorously and at the same time allows for more objective evaluation of tentative new theoretical ideas.

5. References

A. Bies, M. Ferguson, K. Katz, and R. MacIntyre. 1995. Bracketing Guidelines for Treebank II Style Penn Treebank Project.

L. Haegeman. 1994. *Introduction to Government and Binding Theory*. Blackwell Publishing.

J. Hockenmaier and M. Steedman. 2007. CCGbank: A Corpus of CCG Derivations and Dependency Structures Extracted from the Penn Treebank. *Computational Linguistics*, 33(3).

M. Honnibal and J. R. Curran. 2005. Creating a Systemic Functional Grammar Corpus from the Penn Treebank. *ACL 2007*, 100:89–96.

M. Johnson. 1998. PCFG models of linguistic tree representations. *Computational Linguistics*, 24(4):613–632.

D. Klein and C. D. Manning. 2003. Accurate unlexicalized parsing. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430.

M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. 1994. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

A. Mengel and W. Lezius. 2000. An XML-based representation format for syntactically annotated corpora. *Proceedings of the Second International Conference on Language Resources and Engineering (LREC)*, 1:121–126.

S. Oepen, D. Flickinger, K. Toutanova, and C. D. Manning. 2004. LinGO Redwoods. *Research on Language & Computation*, 2(4):575–596.

Y. Samuelsson and M. Volk. 2005. Presentation and representation of parallel treebanks. *Proc. of the Treebank-Workshop at Nodalida*.

P. Tapanainen and T. Järvinen. 1997. A non-projective dependency parser. *Proceedings of the 5th Conference on Applied Natural Language Processing*, pages 64–71.