Porting an Ancient Greek and Latin Treebank

John Lee[*], **Dag Haug**[†]

[*]Department of Chinese, Translation and Linguistics, City University of Hong Kong
jsylee@cityu.edu.hk

[†]IFIKK, University of Oslo
daghaug@ifikk.uio.no

Abstract

We have recently converted a dependency treebank, consisting of ancient Greek and Latin texts, from one annotation scheme to another that was independently designed. This paper makes two observations about this conversion process. First, we show that, despite significant surface differences between the two treebanks, a number of straightforward transformation rules yield a substantial level of compatibility between them, giving evidence for their sound design and high quality of annotation. Second, we analyze some linguistic annotations that require further disambiguation, proposing some simple yet effective machine learning methods.

1. Introduction

A large number of treebanks are now available to support linguistic analyses. Reflecting different grammar formalisms and research agendas, these treebanks follow a variety of annotation schemes. Some well-known examples include the constituent-based Penn Treebank (Marcus et al., 1993), the dependency-based Prague Dependency Treebank (Hajičová et al., 1999), and the hybrid TIGER Treebank (Brants et al., 2002), not to mention variations in the tagsets and other conventions within these traditions.

Despite the differences among these treebanks, however, there is often a substantial overlap of core linguistic information. Therefore, rather than annotating an identical text from scratch, researchers have been developing algorithms for automatically converting from one formalism to another (Forst, 2003; Hockenmaier and Steedman, 2007), and for inserting new information to an existing treebank (Kingsbury and Palmer, 2002; Miltsakaki et al., 2004). By reducing duplication of manual effort, automatic conversion serves as a cost-effective means of rapidly increasing the size of a treebank.

2. Research Questions

In a similar spirit, we recently converted dependency trees in ancient Greek and Latin from one treebank to another. Although both are inspired by the Prague Dependency Treebank, these two treebanks follow significantly different annotation schemes. Among sentences they have in common, agreement in unlabeled dependency arcs is around 70%; the labels in this subset of arcs agree about 55% of the time. In this paper, we seek to answer two questions:

1. Despite their substantial differences on the surface, how compatible are the dependency annotations in these two treebanks?
2. To what extent can the conversion process be automated?

Our data provide us with a unique perspective for question (1). Almost all previous investigations on inter-annotator agreement have been conducted within a single project (Brants, 2000; Civit et al., 2003), whose human annotators underwent similar training sessions, followed the same annotation guidelines, and used the same tagset and annotation tools. In contrast, we compare two treebanks that were *independently designed and annotated*. We will demonstrate a high degree of equivalence between them (§5.1.), thereby providing evidence not only of their individual consistency and accuracy, but also of the soundness and generality in the design of both. In this respect, this paper is similar to an analysis of word-sense annotations on two sets of independently developed sense classes (Ng et al., 1999). To the best of our knowledge, this paper represents the first such analysis on dependency treebanks.

Even if found to be largely compatible, the two treebanks are still expected to exhibit different linguistic judgments; one treebank might include extra information for certain linguistic phenomena, or encode finer-grained distinctions than the other treebank. In addressing question (2), we will identify two challenging areas in our conversion process (§5.2.), namely the annotation of infinitives and prepositional phrases, and describe some statistical methods to perform disambiguation.

3. Data

The two largest dependency treebanks in ancient Greek and Latin have been developed by the PERSEUS (Bamman and Crane, 2007; Bamman et al., 2009) and PROIEL (Haug and Jøhndal, 2008) projects. The 100K-word PERSEUS treebank draws from a wide variety of Greek and Latin literature. The PROIEL (*Pragmatic Resources of Old Indo-European Languages*) treebank focuses on the New Testament, including its Greek original (100K words) and translations in Latin and other old Indo-European languages. We are interested in automatically extending the PROIEL treebank to extra-biblical texts. Hence, our goal is to automatically convert all dependency trees from PERSEUS, henceforth the "source treebank", to PROIEL, henceforth the "target treebank".

The conversion algorithm has been designed and evaluated on three books that are common to both treebanks: the Latin version of the Book of *Revelation* (7061 words) serves as our development set; selections from the Greek version of *Revelation* (4156 words, henceforth the "Greek test set") and from *The Gallic War* (1116 words, henceforth the "Latin test set") form our test sets. These test sets evaluate the algorithm's ability to generalize from the development set across both language and genre, namely from Latin to Greek, and from apocalyptic literature to historiography.

4. Treebank Comparison

Broadly speaking, from the point of view of the target treebank, the source treebank lacks the following two kinds of annotations:

- In the target treebank, null elements are explicitly inserted as "empty nodes" in the trees, rather than implicitly encoded in the dependency labels, as done in the source. These nodes serve to capture ellipsis of conjunctions and verbs, such as the copula *sit* ("be") in the following sentence:

  (1) *gratia Domini nostri Iesu Christi* ***[sit]*** *cum*
  grace Lord our Jesus Christ **[be]** with
  omnibus
  all
  'The grace of our Lord Jesus Christ be with all'

  The elided copula is represented as ⟨`empty`⟩ in the tree, as shown in Table 1(b).

- When a non-finite verb does not have an overt subject, but rather shares a subject with another verb, this "external subject" is annotated with a "slash"[1] in the target treebank. For instance, in Table 1(c), the participle *echōn* ("having") is linked to its subject *zōa* ("living creatures") not by a dependency arc but by a slash, represented by the dotted arrow.

A few other linguistic elements are annotated in both treebanks, but the meanings are not directly comparable, and thus a mechanical conversion is not always possible:

- More fine-grained distinctions are made in the target labels for certain types of infinitives, conjunctions, and objects. For example, an "object" label in the source may correspond to an oblique, a direct object, or complement in the target. Table 1(a) illustrates such a case for pronouns:

  (2) *apocalypsis Iesu Christi quam dedit Deus*
  revelation Jesus Christ which gave God
  illi *...*
  him ...
  'The revelation of Jesus Christ, which God gave him ...'

  The word *illi* is considered an "object" (`OBJ`) in the source but more precisely as an "oblique" (`OBL`) in the target. The case for infinitives will be pursued further in §5.2.1..

- A prepositional phrase (PP) may function as an argument or an adjunct. This distinction remains an active area in linguistics research (Kay, 2005). In computational linguistics, automatic classification of argumenthood is still not highly accurate, even in resource-rich languages such as English (Merlo and Ferrer, 2006). Since decisions on argumenthood can be rather subjective, researchers often develop heuristics to suit their own purposes (Kinyon and Prolo, 2002; Hockenmaier and Steedman, 2007). This situation is reflected in our data — the "dividing line" between argument and adjunct seems to be drawn differently in the two treebanks. This issue will be pursued further in §5.2.2..

- Dependency structures involving appositions, participles, coordinations and conjunctions are different in the two treebanks. Table 1(c) gives an example on participles:

  (3) *kai ta tessera zōa ...* ***echōn***
  and the four living creatures ... **having**
  ana pterugas hex ... gemousin
  upon wings six ... covered with
  ophthalmōn
  eyes
  'Each of the four living creatures had six wings ... and was covered with eyes ...'

  In the target treebank, the head of the participle *echōn* is the main verb *gemousin*, rather than the subject of the participle, *zōa*.

5. Approach and Results

In view of these differences, the conversion process, especially in the direction pursuing in this paper (i.e., from PERSEUS to PROIEL), necessitates both mechanical changes and more subtle disambiguations. Our approach thus consists of two phases. In the first phase (§5.1.), aiming at the more systematic differences, a number of deterministic transformation rules are applied on the source dependency trees. In the second phase (§5.2.), a statistical approach is taken to make distinctions involving infinitives and prepositional phrases, both of which have significant influence on Latin and Greek syntax.

[1] The reader is directed to (Haug and Jøhndal, 2008) for a detailed discussion on the slash notation.

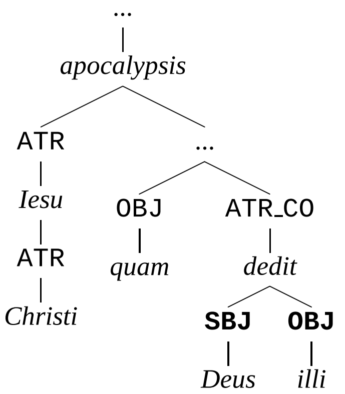
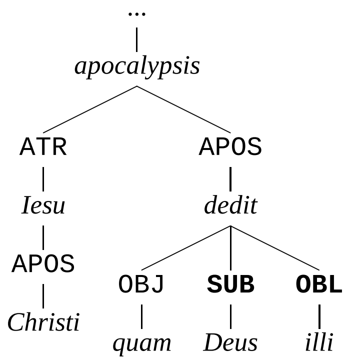
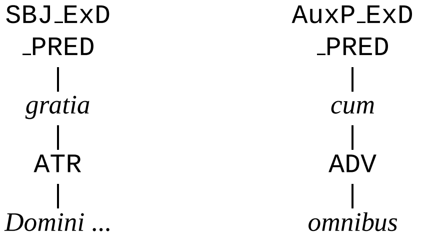
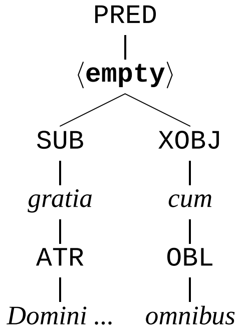
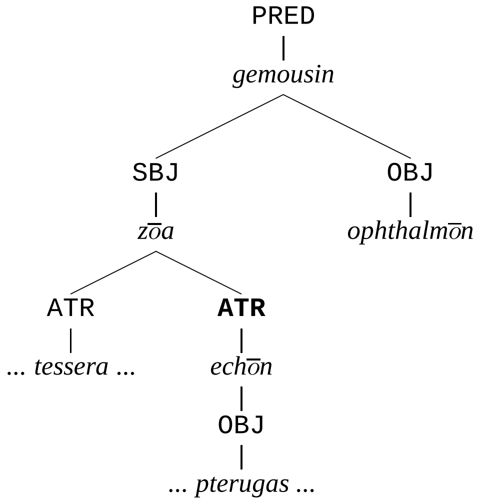
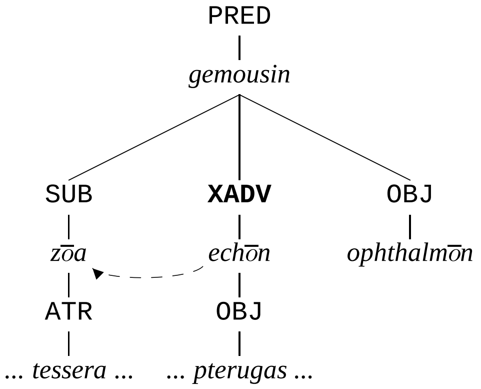
Source Dependency Tree	Target Dependency Tree
(a) *Apocalypsis Iesu Christi quam dedit illi Deus ...* (Revelation 1:1)	
 ... *apocalypsis* ATR ... *Iesu* OBJ ATR_CO ATR *quam* *dedit* *Christi* **SBJ** **OBJ** *Deus* *illi*	 ... *apocalypsis* ATR APOS *Iesu* *dedit* APOS OBJ **SUB** **OBL** *Christi* *quam* *Deus* *illi*
(b) *gratia Domini nostri Iesu Christi cum omnibus* (Revelation 22:21)	
 SBJ_ExD _PRED *gratia* ATR *Domini ...* AuxP_ExD _PRED *cum* ADV *omnibus*	 PRED ⟨**empty**⟩ SUB XOBJ *gratia* *cum* ATR OBL *Domini ...* *omnibus*
(c) *kai ta tessera zōa ... echōn ana pterugas hex ... gemousin ophthalmōn* (Revelation 4:8)	
 PRED *gemousin* SBJ OBJ *zōa* *ophthalmōn* ATR **ATR** *... tessera ...* *echōn* OBJ *... pterugas ...*	 PRED *gemousin* SUB **XADV** OBJ *zōa* *echōn* *ophthalmōn* ATR OBJ *... tessera ...* *... pterugas ...*

Table 1: Example pairs of dependency trees in the source and target treebanks, from which transformation rules are derived:
- In (a), the "subject" label is changed from SBJ to SUB, while the "object" label (OBJ) is re-classified as "oblique" (OBL). The latter change takes morphological information into account, performed only when the word is dominated by a verb and does not have the nominative or accusative case.
- In (b), an empty node is added to represent the elided copula.
- In (c), the main verb (*gemousin*), rather than the subject (*zōa*), becomes the head of the participle (*echōn*); also, a slash is added from the participle to the subject.

5.1. Transformation Rules

Based on pairs of source and target trees in the development set, we derived a dozen of subtree transformation procedures, some of which are illustrated in Table 1. They insert additional annotations, such as null elements and slashes; they also re-structure dependencies involving appositions, coordinations and conjunctions; finally, they re-annotate a number of underspecified source labels that

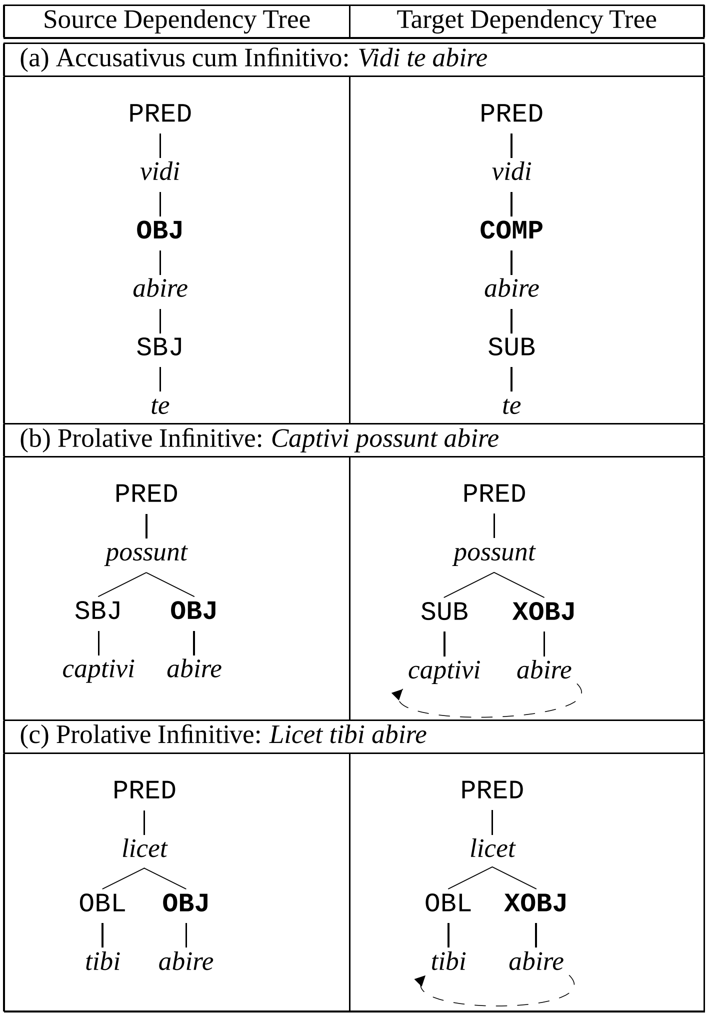


Table 2: Annotation examples of the *Accusativus cum Infinitivo* (AcI) and the prolative infinitive. Infinitives of either type are always labeled `OBJ` in the source, and need to be converted in the target to either `COMP` ("complement") for AcI, or `XOBJ` ("external object") for prolative infinitive. Further, depending on the context, slashes might need to be added from the infinitive to an external subject, as in (b), or to an oblique, as in (c).

are more readily predictable.

These procedures resulted in 85.2% agreement in unlabeled arc dependencies for the Latin test set and 81.0% for the Greek. Within this subset of arcs, label agreement is 84.6% and 91.9% respectively (excluding those involving infinitives and prepositional phrases, which will be treated in §5.2.). Given that the two treebanks have been independently developed, these figures provide good evidence for the sound design of both treebanks. They can also be viewed as an estimate of the ceiling of human annotation accuracy for languages with a higher degree of morphological ambiguity.

A frequent category of disagreement is between the labels "attribute" (`ATR`) and "apposition" (`APOS`). Two instances can be seen in Table 1(a). The source annotator considers *Christi* to be an attribute of the person *Iesu*, whereas the target annotator deems it to be an apposition. The same disagreement occurs for the phrase *dedit illi Deus*, on whether it is an attribute of, or an apposition to, the *apocalypsis*.

5.2. Statistical Disambiguation

While the rules described in §5.1. are effective in most aspects of the conversion, it is difficult to manually derive rules to annotate certain kinds of infinitives and prepositional phrases. We therefore took a machine-learning approach, using the rest of the target treebank as training material.

5.2.1. Infinitives

When dominated by another verb, an infinitive is annotated in the target treebank as either *Accusativus cum Infinitivo* (AcI) or prolative infinitive (Pinkster, 1990). This distinction is critical in identifying verb subcategorization frames, which are in turn important in the comparative study of Indo-European languages, the key goal of the PROIEL project.

An AcI introduces a complement clause with its own subject in the accusative case. For example, in (4), the accusative *te* serves as the subject of the complement clause headed by the infinitive *abire*, meaning 'I saw *that you left*',

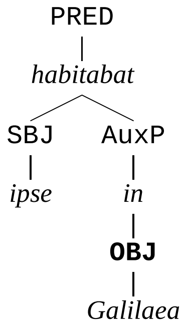
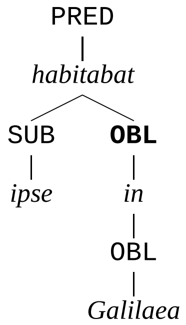
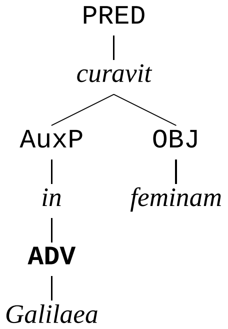
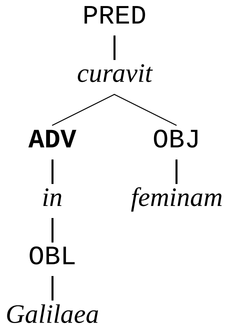
Source Dependency Tree	Target Dependency Tree
(a) Argument PP: *Ipse habitabat in Galilaea*	
 PRED habitabat SBJ AuxP ipse in **OBJ** Galilaea	 PRED habitabat SUB **OBL** ipse in OBL Galilaea
(b) Adjunct PP: *In Galilaea feminam curavit*	
 PRED curavit AuxP OBJ in feminam **ADV** Galilaea	 PRED curavit **ADV** OBJ in feminam OBL Galilaea

Table 3: Annotation examples of the argument-adjunct distinction in prepositional phrases. In both treebanks, the label `ADV` stands for adjunct; the label `OBJ` stands for argument in the source, but `OBL` is used in the target. However, in the source, the information is encoded in the dependency arc from the complement, while in the target, it is encoded in the arc from the preposition. In both example pairs, the argumenthood labels happen to be the same, but in general, a binary classification must be performed to determine the appropriate target label.

rather than as the object of *vidi*, which would mean 'I saw you leave'.

(4) *Vidi te* ***abire***
I saw you **to leave**
'I saw that you left'

(5) *Captivi possunt* ***abire***
captives can **to leave**
'The captives may leave'

(6) *Licet tibi* ***abire***
is permitted for you **to leave**
'You are permitted to leave'

In contrast, a prolative infinitive is just an infinitival complement with an external subject. In example (5), the infinitive does not have its own subject, but shares the subject of the main verb *possunt*; in (6), the infinitive has a dative dependent on the main verb as its subject.

As illustrated in Table 2, the source treebank makes no distinction between AcI and the prolative infinitive; an infinitive of either type is labeled `OBJ`. However, in the target, an AcI is assigned `COMP` and a prolative infinitive is assigned `XOBJ`. A binary classification must therefore be performed.

A simple baseline is to assign AcI whenever a subject in the accusative case is found, as in the case of Table 2(a), and otherwise default to prolative infinitive, which occurs more frequently than AcI. This baseline yielded 77.4% accuracy in the Latin test set and 86.7% in the Greek.

Unfortunately, such accusative subjects are frequently elided when they can be inferred from the context; hence, example (4) may be re-written as "*Vidi abire*", and (6) as "*Licet abire*", in which case the resulting trees would be indistinguishable from each other. So, one cannot rely on the presence of overt accusatives alone to make the distinction. Instead, for each verb, we consider all instances in the training set where it dominates an infinitive without an accusative subject, and compare the relative frequencies of the `COMP` label against `XOBJ`. When testing, we apply the more frequent label.

On the Latin test set, this approach resulted in an absolute improvement of 6.5% over the baseline. However, on the Greek test set, which has a higher baseline, it failed to achieve any improvement. Each of these test sets consists of only about 30 instances; more data would be desirable to fully evaluate this approach.

5.2.2. Prepositional Phrase Argumenthood

The argument-adjunct distinction in prepositional phrases (PPs) is another annotation task that is not amenable to hand-crafted rules. Detailed discussions on this distinction can be found in the literature such as (Kay, 2005); due to space constraints, we illustrate with only a simple example:

(7) *Ipse* *habitabat* ***in Galilaea***
he himself lived **in Galilee**
'He himself lived in Galilee'

(8) ***In Galilaea*** *feminam curavit*
in Galilee woman he healed
'In Galilee he healed a woman'

In (7), the PP is closely tied to the verb *habitabat*, which demands a complement of place, so the PP is an argument, and is annotated in the way shown in Table 3(a). In (8), the PP just gives information about where the event *curavit* took place and is considered an adjunct, as shown in Table 3(b). A trivial baseline of simply using the source label yielded 68.0% accuracy in the Latin test set and 55.2% in the Greek.

In general, the source treebank identifies fewer PPs as arguments, and in these cases the target treebank often concurs. We focus our effort, therefore, on classifying the adjunct PPs in the source treebank, using the nearest-neighbor framework. This framework has been shown to perform well on a variety of benchmark tasks in natural language processing (Daelemans et al., 1999) and has been applied, in particular, to the related task of preposition generation (Lee and Knutsson, 2008).

Four features are extracted from each PP in the training set — its head verb, preposition, complement, and case of the complement. When testing, the same features are extracted from the PP, and the algorithm looks for PPs in the training set with identical feature values. The majority label (i.e., "argument" or "adjunct") among these "nearest neighbors" is returned. When no such PP exists in the training set, the algorithm backs off to an overlap of three out of the four features, and continues to back off if necessary. This strategy yielded absolute improvements of 6.8% and 16.1% in the Latin and Greek test sets, respectively, over the baseline.

6. Conclusions

We have described the conversion process of an ancient Greek and Latin dependency treebank, using a combination of transformation rules and statistical methods. Overall, in the Latin test set, 85.2% of the unlabeled dependency arcs agree, and within this subset of arcs, 83.5% of their labels agree. The respective figures for the Greek test set are 81.0% and 90.4%. Human post-processing will still be needed, but the automatic conversion should substantially reduce the time and effort required.

We draw two conclusions. First, the substantial level of compatibility between the two treebanks (§5.1.) gives compelling evidence for their sound design and high quality of annotation. Second, for annotating infinitives and prepositional phrases, the machine learning approaches described in §5.2. show promising results. Their simplicity makes them potentially applicable to other low-resource languages with modest-sized treebanks.

7. Acknowledgments

We thank the Perseus Project, Tufts University, for providing the Greek test set. The first author gratefully acknowledges the support of the Faculty of Humanities at the University of Oslo, where he conducted part of this research.

8. References

David Bamman and Gregory Crane. 2007. The latin dependency treebank in a cultural heritage digital library. In *Proc. ACL Workshop on Language Technology for Cultural Heritage Data*, Prague, Czech Republic.

David Bamman, Francesco Mambrini, and Gregory Crane. 2009. An ownership model of annotation: The ancient greek dependency treebank. In *Proc. 8th International Workshop on Treebanks and Linguistic Theories (TLT)*, Milan, Italy.

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The tiger treebank. In *Proc. Workshop on Treebanks and Linguistic Theories (TLT)*, Sozopol, Bulgaria.

Thorsten Brants. 2000. Inter-annotator agreement for a german newspaper corpus. In *Proc. LREC*, Athens, Greece.

Montserrat Civit, Alicia Ageno, Borja Navarro, Nuria Bufi, and M. A. Martí. 2003. Qualitative and quantitative analysis of annotators' agreement in the development of cast3lb. In *Proc. 2nd Workshop on Treebanks and Linguistic Theories (TLT)*, Växjö, Sweden.

Walter Daelemans, Antal van den Bosch, and Jakub Zavrel. 1999. Forgetting exceptions is harmful in language learning. *Machine Learning*, 34:11–41.

Martin Forst. 2003. Treebank conversion establishing a testsuite for a broad-coverage lfg from the tiger treebank. In *Proc. EACL Workshop on Linguistically Interpreted Corpora*, Budapest, Hungary.

Eva Hajičová, Zdeněk Kirschner, and Petr Sgall. 1999. A manual for analytic layer annotation of the prague dependency treebank (english translation). Technical report, UFAL MFF UK, Prague, Czech Republic.

Dag Haug and Marius Jøhndal. 2008. Creating a parallel treebank of the old indo-european bible translations. In *Proc. LREC Workshop on Language Technology for Cultural Heritage Data*.

Julia Hockenmaier and Mark Steedman. 2007. Ccgbank: A corpus of ccg derivations and dependency structures extracted from the penn treebank. *Computational Linguistics*, 33(3):355–396.

Paul Kay, 2005. *Grammatical Constructions: Back to the Roots*, chapter Argument Structure Constructions and the Argument-Adjunct Distinction, pages 71–98. John Benjamins, Amsterdam, Netherlands.

Paul Kingsbury and Martha Palmer. 2002. From treebank to propbank. In *Proc. LREC*, Las Palmas, Spain.

Alexandra Kinyon and Carlos Prolo. 2002. Identifying verb arguments and their syntactic function in the penn treebank. In *Proc. LREC*, Las Palmas, Spain.

John Lee and Ola Knutsson. 2008. The role of pp attachment in preposition generation. In *Proc. 9th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, Haifa, Israel.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19:313–330.

Paola Merlo and Eva Esteve Ferrer. 2006. The notion of argument in pp attachment. *Computational Linguistics*, 32(2).

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2004. The penn discourse treebank. In *Proc. LREC*, Lisbon, Portugal.

Hwee Tou Ng, Chung Yong Lim, and Shou King Foo. 1999. A case study on inter-annotator agreement for word sense disambiguation. In *Proc. ACL SIGLEX Workshop on Standardizing Lexical Resources*, College Park, MD.

Harm Pinkster. 1990. *Latin Syntax and Semantics*. Routledge, London, England.