Dependency Parsing using Prosody Markers from a Parallel Text

John Lee

Department of Chinese, Translation and Linguistics
City University of Hong Kong
E-mail: jsylee@cityu.edu.hk

Abstract

This paper describes a method to automatically generate dependency trees for ancient Greek sentences by exploiting prosodic annotation in a Hebrew parallel text. The head selection accuracy of the resulting trees, at close to 80%, is significantly higher than what standard statistical parsers might be expected to produce, for a resource-poor language such as ancient Greek. Our evaluation suggests that prosodic markers can be reliable indicators of syntactic structures.

1 Introduction

Increasingly, researchers in digital humanities are exploiting statistical techniques in the study of ancient languages, including decipherment [1, 2], morphology [3], and syntax [4, 5]. Data-driven syntactic analysis requires large treebanks, which are labour intensive and time consuming to create, especially so when the language in question no longer has any native speakers.

Ancient Greek is an important vehicle of human civilization, but relatively little syntactic annotation has been performed on its literature. Currently, the largest treebanks, both manually crafted, are the 200K-word Perseus Greek Dependency Treebank [6] and the 100K-word *Pragmatic Resources of Old Indo-European Languages* (PROIEL) [7]. An enormous amount of historically significant texts await analysis — the *Thesaurus Linguae Graecae* alone has more than 105 million words in its electronic collection. Although statistical parsers can be trained on these existing treebanks, their performance is unlikely to be adequate. Parsing accuracy has reached the nineties for English, but it is significantly lower for resource-poor languages [8], and lower still for classical Latin, a language with comparable characteristics and digital resources as ancient Greek: the state-of-the-art accuracy is about 54% [4][1].

[1] Accuracy for medieval Latin, however, is higher at about 80% [5].

This paper describes a method to automatically derive dependency trees in ancient Greek, by exploiting prosodic markers from a word-aligned parallel text in ancient Hebrew. The contribution of this paper is two-fold. First, the resulting treebank of the *Septuagint*, the Greek text with which we are concerned, contains 0.6 million words, doubling the size of the existing treebanks. Second, our evaluation shows that prosodic information can help produce parse trees of significantly higher accuracy than would be expected of those produced by a statistical parser, if trained on currently available resources.

2 Research Background

2.1 Previous Work

There has been much research on transferring the syntactic analysis of a resource-rich language, L_1, to a resource-poor language, L_2, given sufficient parallel text for the two languages in question. A popular approach is *syntactic projection* [9, 10], founded on the Direct Correspondence Assumption for syntactic structures. L_1 parse trees and L_1-L_2 word alignments are first acquired, either manually or automatically; dependencies between words in L_1 are then projected onto their L_2 counterparts, possibly followed by some local transformations as required by the linguistic peculiarities of L_2. In [10], using gold-standard word alignments and L_1 dependency trees, syntactic projection yielded unlabelled attachment F-scores of 70.3% for English-to-Spanish, and 67.3% for English-to-Chinese, a more divergent pair of languages.

In *bilingual parsing* [11, 12], a unified model performs joint inference for the best L_1 and L_2 parse trees, as well as the word alignments. This approach works well when there is noise in the L_1 parse trees, since it is able to find the combination of parse trees and alignments that are collectively more likely. It has been shown to improve Korean parsing when coupled with English [12].

With ancient Hebrew as L_1 and ancient Greek L_2, our work is analogous to syntactic projection, but with one crucial difference — we do not, in contrast to [10], have the luxury of a high-performing L_1 parser. Instead of L_1 dependency trees, our prior information will be prosodic annotation known as cantillation marks, for which we now provide some background.

2.2 Cantillation Marks

In order to facilitate public chanting of the Hebrew Bible, a group of scholars called the Masoretes added special symbols, called *cantillation marks*, to the text between the 7th and 10th century CE. These marks, written above or beneath a word, may be considered to be very fine-grained punctuation marks. They fall into one of two categories. Simply put, a word bearing a *disjunctive mark* suggests that a prosodic boundary separates it from the following word; a word bearing a *conjunctive mark*, in contrast, indicates that there is no such prosodic boundary.

Category	Names (listed from strongest to weakest)
Disjunctive	passuq [·], atnah [;], segolta, little zaqef, great zaqef, tifha [,], revia, zarqa, pashta, yetiv, tevir, geresh, pazer, great pazer, great telisha
Conjunctive	maqqef [≡], munah [=], mehuppakh, merekha-khefula [-], darga, azla, little telisha, galgal

Table 1: Cantillation marks are either *disjunctive* or *conjunctive*. They are listed above in descending order of strength [13]. The symbols in square brackets are shorthands used in the rest of the paper, and do not represent their actual shapes.

Hebrew	*wy'mr* [-]	*'lhym* [,]	*yhy* [=]	*'wr* [;]	*wyhy* [≡]	*'wr* [·]
English	and said	the God	let be	light	and became	light

Table 2: The original Hebrew words from Genesis 1:3, *And God said, "Let there be light", and there was light* (translation from Jewish Publication Society). Cantillation marks are shown in square brackets. See §2.2 for a discussion on how the marks help disambiguate this sentence.

These boundaries, however, should be understood in a relative sense, since these marks are organized in a complex hierarchy according to their levels of "strength"[2], as listed in Table 1 [13].

Cantillation marks can help interpret a sentence. Consider, for example, the Hebrew sentence in Table 2. Based on the words alone, it may be read as:

> *And God said, "Let there be light", and there was light.*

but also possibly as:

> *And God said, "Let there be light and there was light."*

However, the cantillation marks strongly suggest the first interpretation. The disjunctive marker *atnah* at the first occurrence of the word "light" is stronger than the *tifha* at the word "God"; hence the pause following the former should be more substantial than the latter.

Correspondence has been demonstrated between clause structures and tone units in English [14]. Cantillation marks have also been hypothesized to correlate with units of meaning and syntactic phrases in Hebrew [15], although no formal evaluation has been reported. In this paper, we use these marks to infer syntactic dependencies in an ancient Greek parallel text, namely the *Septuagint*, and evaluate the accuracy of the resulting dependency trees.

[2]Among the conjunctive marks, the *maqqef* is the strongest. The differences among the rest are of a musical nature only and are ignored for our purpose.

Hebrew	*wy'mr* [-]	*'lhym* [,]	*yhy* [=]	*'wr* [;]	*wyhy* [≡]	*'wr* [·]
Greek	*kai eipen*	*ho theos*	*genēthētō*	*phōs*	*kai egeneto*	*phōs*

Table 3: Hebrew-to-Greek word alignments associate Hebrew words to zero, one or more Greek words; these will be referred to as a "Greek chunk".

POS combination	Frequency
Noun	18%
Verb	13%
Article + Noun	12%
Conjunction + Verb	8%
Pronoun	6%

Table 4: The most common parts-of-speech combinations for the Greek chunks extracted from the word alignments (see Table 3). For example, the chunk "*ho theos*" has the combination "Article + Noun".

3 Approach

In addition to the Hebrew text annotated with prosodic marks, we have at our dispoal Hebrew-to-Greek word alignments, some samples of which are shown in Table 3. In general the alignments are many-to-many, but most of the time one Hebrew word is associated with zero, one or two Greek words, which will be referred to as a "Greek chunk" in the rest of the paper. These chunks have an average length of 1.6 words, and will serve as the starting blocks of the derivation process of the dependency tree. The most frequent part-of-speech (POS) combinations for the chunks are listed in Table 4.

Following [7] and [18], we adopt the dependency tree [16] as the target syntactic representation. Our approach consists of two main phases. After some preliminary steps (§3.1), the first phase (§3.2) constructs dependency subtrees for each Greek chunk; the second phase (§3.3) merges these subtrees, two at a time, in an order determined by the Hebrew cantillation marks. The division of labor between these two phases is similar to that between the chunker and attacher in [19].

3.1 Preliminary Steps

Before parsing can begin, two preliminary steps must be taken to deal with incomplete word alignment and different word orders.

3.1.1 Insertion

In about 40% of the sentences, one or more Greek words are not aligned with any Hebrew ones. In some cases, they reflect genuine differences in the content, due

to textual variations between the Hebrew *Vorlage* and our particular *Septuagint* version. But far more often, non-alignment is caused by differences in syntax. For example, the Hebrew noun construct chain "*yšby* ⟨place⟩" (literally, "dwellers-of ⟨place⟩") is often rendered in Greek as the participial form of "dwell", followed by the preposition "*en*" ("in") and ⟨place⟩, such as "*katoikountas en tais polesin*" ("those dwelling in the cities", Genesis 19:25).

In other instances, stylistic considerations are responsible for the non-alignment. For example, the Greek verb-to-be "*ēn*" is used in the phrase "*lithos de ēn megas*" ("*the stone ... was large*", Genesis 29:2), where the original Hebrew has none.

These non-aligned Greek words may form their own chunk. Alternatively, they may be amalgamated with the chunk to its left; a common example is the joining of a postpositive particle, such as "*oun*" ("therefore"), to the preceding verb. Lastly, they may join the chunk on their right, as would be appropriate for the preposition "*en*" heading a prepositional phrase, such as in the example above. Among these three options, we choose the one that would yield a chunk whose POS combination has the highest frequency count, based on statistics computed from the rest of the corpus.

3.1.2 Re-ordering

The *Septuagint* is a highly literalistic translation; the relatively free word order in the Greek language allowed translators to largely conserve the word order in the original. Indeed, ignoring insertions and deletions, the Greek word order exactly matches the Hebrew word order in 94% of the verses. This high percentage facilitates the preservation of prosodic boundaries from the Hebrew to the Greek.

Two systematic exceptions involve particles and numbers. One is the use of postpositive particles such as "*gar*" and "*de*", which must be placed after the first word, to render the Hebrew sentence-initial "*ky*". For noun phrases involving a number and the word "year" or "day", the number usually comes first in Hebrew but comes last in Greek. Some other differences in word order are caused by the relative positions of verbs and their direct or indirect objects.

3.2 Parsing Greek Chunks

After addressing issues with word alignments and word order, we can now derive dependency trees for the Greek chunks. A dozen rules, each targeting chunks of a specific POS combination, were written to assign dependency relationships within the chunk, using the guidelines for [7]. Table 5 shows a rule for chunks of the type "Article + Noun" being applied on the chunk "*ho theos*"; the noun "*theos*" is annotated as the head of the article "*ho*", with the relation AUX.

These rules make use of not only the POS combination but also morphological information. Consider the chunks "*eneteilamēn soi*" ("I commanded you") and

Greek chunk	*kai eipen* [-] “and said”	*ho theos* [,] “the God”	*kai egeneto* [≡] “and became”
Subtree	[-] *kai* | *eipen* PRED	[,] *theos* | *ho* AUX	[≡] *kai* | *egeneto* PRED

Table 5: Derivation of dependency subtrees (see §3.2) for the Greek chunks, taken from the sentence in Table 3. (The chunks of length 1 are omitted.) Note that the Hebrew cantillation marks have been projected onto the chunks.

“*ginōskō egō*” (“I know”); both consist of a verb followed by a personal pronoun. In the first, the pronoun “you” is dative and hence is an indirect object of the verb; in the second, in contrast, the pronoun “I” is nominative and is the subject of the verb.

3.3 Merging Subtrees

For each sentence, the procedure in §3.2 yields a sequence of subtrees, such as those in Table 5; they must then be merged into a single dependency tree. The merging process requires two kinds of decisions: the merge order, which will be determined by the relative strengths of the cantillation marks; and the attachment site, which will be informed by manually derived rules.

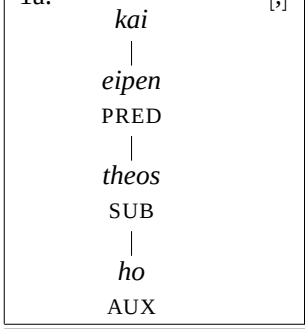
3.3.1 Merge Order

Using Hebrew-to-Greek word alignments, cantillation marks are projected from each Hebrew word to its corresponding Greek chunk and subtree (Table 5). Only the mark on the last word of a chunk is retained; the rest are ignored. Following the analogous treatment of the Hebrew in [17], the Greek subtrees are then merged two at a time. First, in descending order of strength, those with conjunctive marks are merged with their right neighbors (step 1 in Table 6); then, in ascending order of strength, those with disjunctive marks are merged with their right neighbors (steps 2 and 3 in Table 6).

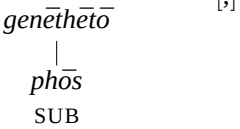
To ensure each Greek chunk has one cantillation mark, two issues need to be resolved. First, when a new chunk is inserted via the insertion step (§3.1.1), its cantillation mark is predicted using an *n*-gram model. We trained a trigram model on the existing chunks, treating a chunk’s POS combination as the “word” and its cantillation mark as the “tag”. Also, if a Hebrew word has a left neighbor which is unaligned but has a stronger disjunctive mark, it will project this stronger mark instead of its own, so as to preserve the prosodic boundary.

Step 1: Merge subtrees linked by conjunctive cantillation marks

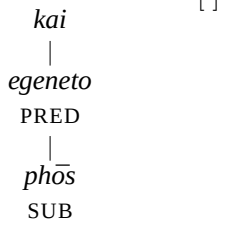
1a. [,]



1b. [;]

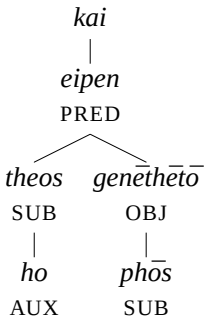


1c. [·]

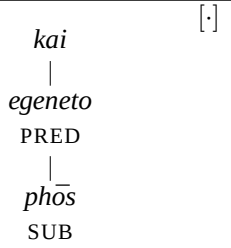


Step 2: Merge subtrees linked by weaker disjunctive cantillation marks

2a. [;]



2b. [·]



Step 3: Merge subtrees linked by stronger disjunctive cantillation marks

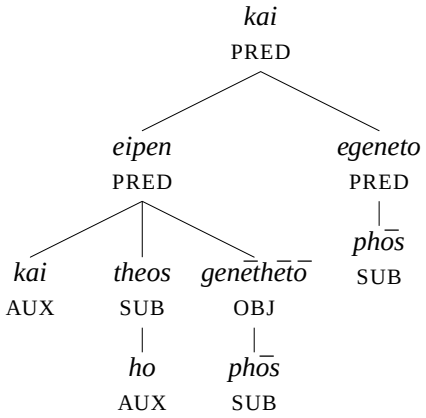


Table 6: The subtree merging process (§3.3) for Genesis 1:3. Step 1 shows the result of merging three pairs of subtrees (Tables 3 and 5) that are connected by conjunctive marks (namely the *merekha*, *munah* and *maqqef*). Step 2 faces two options: merge 1b ("Let there be light") and 1c ("And there was light") first, or merge 1a ("And God said") and 1b first. The first option would have yielded a tree with the interpretation *And God said, "Let there be light and there was light"*. See a discussion in §2.2 on how the stronger disjunctive mark between 1b and 1c rules out this option. Finally, 2a and 2b ar[illegible] as coordinated predicates, resulting in the final tree in step 3.

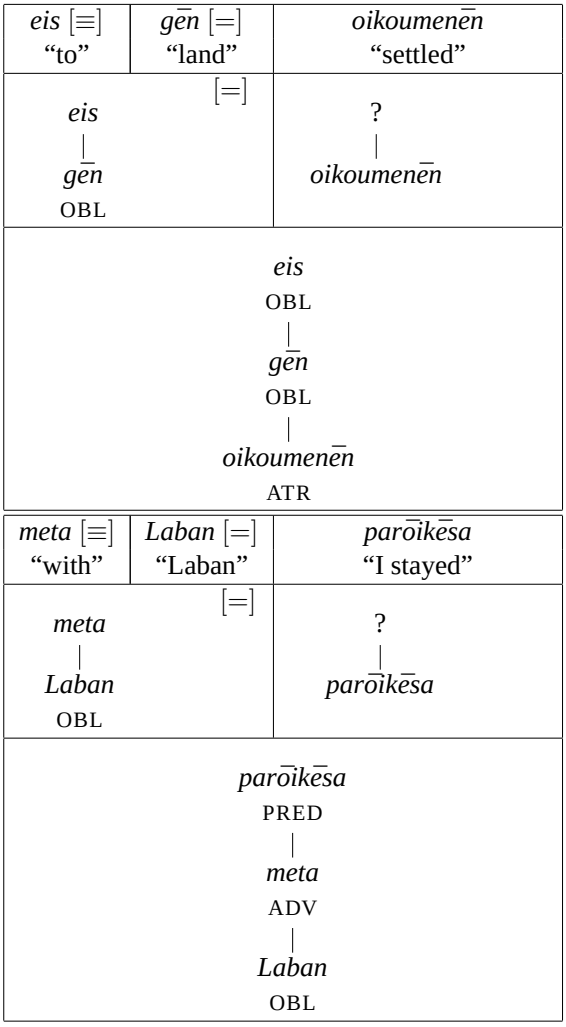


Table 7: Examples of subtree merging involving function words, as discussed in §3.3.2. Both are preposition-noun-verb trigrams with identical cantillation marks, but their merged trees are completely different. The verb in the top example (Exodus 16:35) is a participle which forms part of a long noun phrase, whereas the verb in the bottom (Genesis 32:5) is finite and becomes the root of the new tree. Morphological analysis of the verbs is indispensable for correct parsing.

3.3.2 Root Attachment Site

Having specified the order, we now turn our attention to how the dependency subtrees are merged. The most straightforward manner is to assign the root word of one subtree as the head of the root of the other. For example, in Genesis 32:5 in Table 7, the verb "*parōikēsa*" is assigned as head of the preposition "*meta*". A verb-object pair would be treated likewise. This decision is made according to the POS of the root words, expressed through a dozen of deterministic rules for each POS pair. Table 6 continues with our running example, completing the entire merging process.

Relative clauses When the prosodic structure differs from the syntactic structure, the appropriate attachment site may not be the root. One such case occurs with a relative clause, whose root is dependent on its antecedent noun; this noun is not necessarily the root of the other subtree. For example, in Genesis 3:3, there are two chunks "*apo de karpou tou xulou*" ("from the fruit of the tree") and "*ho estin en mesō tou paradeisou*" ("that is in the middle of the garden"). The relative pronoun "*ho*" signals that the root "*estin*" ("is") must be attached to its antecedant noun. The algorithm searches within the first subtree in post-order, to find a noun — in this case, "*xulou*" ("tree") — that agrees with the relative pronoun with respect to gender and number. Hence, "*xulou*", rather than the head "*apo*" ("from"), becomes the head of "*estin*".

Function words and noun phrases A systematic disagreement between the cantillation marks and phrase boundaries occurs when a long noun phrase depends on a function word, such as a conjunction or a preposition. Consider the phrase "*eis* [≡] *gēn* [=] *oikoumenēn*" ("to a land that was settled", illustrated in Table 7). The strongest conjunctive mark, the *maqqef*, binds the preposition "*eis*" ("to") to "*gēn*" ("a land"), the first word of the noun phrase; it thus tears apart the two-word NP "a land [=] that was settled", which is held together by a weaker conjunctive mark, the *munah*. This makes sense prosodically, since a pause is needed in the middle of a long NP, as suggested in the discussion of a similar phenomenon in Hebrew in [15]. The procedure described above would have produced an incorrect tree; instead, *oikoumenēn* should be dependent on the non-root *gēn*.

Morphological analysis is necessary to decide whether this kind of adjustment is warranted. Consider another phrase with a preposition-noun-verb sequence, "*meta* [≡] *Laban* [=] *parōikēsa*" ("with Laban I have been staying", illustrated in Table 7). Its surface structure and cantillation marks are indistinguishable from the last example. However, the finiteness of the verb suggests that the preposition should be dependent on it.

4 Evaluation

4.1 Data

The book of Genesis is used as the development set to design the rules for parsing Greek chunks (§3.2) and merging subtrees (§3.3). Three poetic books in the *Septuagint*, namely Job, Psalms and Proverbs, use a system of cantillation marks that differ from the one presented in §2.2. They were therefore excluded from the evaluation. The cantillation marks are extracted from the corpus described in [15]. The Hebrew-to-Greek word alignments and the morphologically analyzed corpus of the *Septuagint* are compiled by the Center for Computer Analysis of Texts at the University of Pennsylvania.

There is no existing treebank for the *Septuagint*. Fortunately, many of its verses appear also in the Greek New Testament, much of which has been analyzed in the PROIEL dependency treebank [7]. Some quotations diverge slightly from the original; to automate the creation of the gold-standard trees, we adopted the simple criterion of including all fragments of at least five consecutive words that are quoted *verbatim*. The gold-standard trees[3] for these fragments were extracted from PROIEL. After these filtering steps, there were altogether 995 words for evaluation.

4.2 Result

The unlabeled attachment score is 79.4%. A comparison with [10], a related work in syntactic projection, is difficult due to different language pairs and text genre; nonetheless, the higher score achieved here provides some evidence that prosodic boundaries are reasonable predictors of syntactic boundaries. A chief problem is the analysis of coordinated phrases, especially those embedded in long sentences. Another source of head selection error is the attachment of relative pronouns, as described in §3.3.2, as well as participles, when it can be modifying one of multiple nouns or verbs.

Among words with correctly selected heads, 88.5% are assigned the correct dependency label, yielding an overall labeled attachment score of 70.6%. This level of accuracy is significantly higher than what statistical parsers might be expected to achieve, if the corresponding score of 54% reported in [4] for classical Latin, a language with similarly limited resources, has any value as a reference point. Among the label errors, the most frequent mistake, constituting more than a fifth of the total, is the confusion between adjunct and argument in prepositional phrases (labeled as ADV and OBL, respectively). This difficulty echoes the findings in [20]; indeed the adjunct-argument distinction remains challenging even for resource-rich languages such as English [21].

[3] If the head of a word is located outside the fragment, the word is excluded, since its head may not be the same in the *Septuagint*.

5 Conclusion and Future Work

We have described a method to automatically create ancient Greek dependency trees by leveraging prosodic annotation in a parallel text in Hebrew. The resulting treebank for the *Septuagint* is a substantial addition to the relative dearth of syntactic data currently available for ancient Greek. It may be expected to help boost the performance of a statistical parser.

This study also provides evidence that cantillation marks are good indicators of syntactic boundaries. Similar techniques can generate treebanks for other languages into which the Hebrew Bible has been translated.

Acknowledgments

We thank the J. Alan Groves Center for Advanced Biblical Research for making available the Hebrew Bible corpus.

References

[1] Lin, Shou-de and Knight, Kevin (2006) Discovering the Linear Writing Order of a Two-Dimensional Ancient Hieroglyphic Script. In *Artificial Intelligence* **170**(4/5):409—421.

[2] Snyder, Benjamin, Barzilay, Regina and Knight, Kevin (2010) A Statistical Model for Lost Language Decipherment. In *Proc. ACL*.

[3] Lee, John (2008) A Nearest-Neighbor Approach to the Automatic Analysis of Ancient Greek Morphology. In *Proc. Conference on Computational Natural Language Learning (CoNLL)*.

[4] Bamman, David and Crane, Gregory (2008) Building a Dynamic Lexicon from a Digital Library. In *Proc. 8th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2008)*.

[5] Passarotti, Marco and Dell'Orletta, Felice (2010) Improvements in Parsing the Index Thomisticus Treebank. In *Proc. LREC*.

[6] Bamman, David, Mambrini, Francesco and Crane, Gregory (2009) An Ownership Model of Annotation: The Ancient Greek Dependency Treebank. In *Proc. TLT*.

[7] Haug, Dag and Jøhndal, Marius (2008) Creating a Parallel Treebank of the Old Indo-European Bible Translations. In *Proc. LREC Workshop on Language Technology for Cultural Heritage Data*.

[8] Nivre, Joakim, Hall, Johan, Kübler, Sandra, McDonald, Ryan, Nilsson, Jens, Riedel, Sebastian and Yuret, Deniz (2007) The CoNLL 2007 Shared Task on Dependency Parsing. In *Proc. EMNLP-CoNLL*.

[9] Yarowsky, David and Ngai, Grace (2001) Inducing Multilingual POS Taggers and NP Bracketers via Robust Projection across Aligned Corpora. In *Proc. NAACL*.

[10] Hwa, Rebecca, Resnik, Philip, Weinberg, Amy, Cabezas, Clara and Kolak, Okan (2005) Bootstrapping Parsers via Syntactic Projection across Parallel Texts. In *Natural Language Engineering* **11**(3):311–325.

[11] Melamed, I. Dan, Satta, Giorgio and Wellington, Benjamin (2004) Generalized Multitext Grammars. In *Proc. ACL*.

[12] Smith, D. A. and Smith, N. A. (2004) Bilingual Parsing with Factored Estimation: Using English to Parse Korean. In *Proc. EMNLP*.

[13] Robinson, David and Levy, Elisabeth (2002) *The Masoretes and the Punctuation of Biblical Hebrew*. British and Foreign Bible Society.

[14] Fang, Alex C., House, Jill and Huckvale, Mark (1998) Investigating the Syntactic Characteristics of English Tone Units. In *Proc. International Conference on Spoken Language Processing (ICSLP)*.

[15] Wu, Andi and Lowery, Kirk (2006) From Prosodic Trees to Syntactic Trees. In *Proc. ACL*.

[16] Mel'čuk, Igor (1988) *Dependency Syntax: Theory and Practice*. State University of New York Press.

[17] Machine Assisted Translation Team (2002) The Masoretes and the Punctuation of Biblical Hebrew. British and Foreign Bible Society. *http://lc.bfbs.org.uk*

[18] Crane, Gregory, Chavez, Robert F., Rydberg-Cox, Jeffrey A., Mahoney, Anne, Milbank, Thomas L. and Smith, David A. (2001) Drudgery and Deep Thought: Designing Digital Libraries for the Humanities. In *Communications of the ACM* **44**(5):175–182.

[19] Abney, Steven. Parsing by Chunks (1991) In Robert Berwick, Steven Abney and Carol Tenny (eds.), *Principle-Based Parsing*, Kluwer Academic Publishers.

[20] Lee, John and Haug, Dag (2010) Porting an Ancient Greek and Latin Treebank. In *Proc. LREC*.

[21] Merlo, Paola and Esteve Ferrer, Eva (2006) The Notion of Argument in PP Attachment. In *Computational Linguistics* **32**(3):341—378.