Producing a Parallel Universal Dependencies Treebank of Ancient Hebrew and Ancient Greek via Cross-Lingual Projection

Daniel G. Swanson, Bryce D. Bussert, Francis M. Tyers
Indiana University, Gateway Seminary, Indiana University
Department of Linguistics, Department of Biblical Studies, Department of Linguistics,
Bloomington, Indiana, Ontario, California, Bloomington, Indiana
dangswan@iu.edu, bussertscholar@gmail.com, ftyers@iu.edu

Abstract

In this paper we present the initial construction of a treebank of Ancient Greek containing portions of the Septuagint, a translation of the Hebrew Scriptures (1576 sentences, 39K tokens, roughly 7% of the total corpus). We construct the treebank by word-aligning and projecting from the parallel text in Ancient Hebrew before automatically correcting systematic syntactic mismatches and manually correcting other errors.

Keywords: treebank, UD, Greek, parallel

1. Introduction

The Hebrew Scriptures are a collection of 39 documents mostly composed in the first millennium BC in Ancient Hebrew, with a few sections in Aramaic. By the first century AD these had all been translated into Ancient Greek in a collection known as the Septuagint.

The Universal Dependencies (UD) project (Nivre et al., 2020) is a collaborative effort to create a collection of treebanks in a single cross-linguistically consistent annotation scheme so as to better facilitate studying syntax in multiple languages.

Parallel treebanks have previously been used to identify and evaluate changes caused by structural linguistic factors and/or interpretive decisions in translation (Eckhoff et al., 2018; Cherney, 2014; Kahn et al., 2009). This aids in source and textual analysis, especially when, as is the case with the Hebrew Bible, a translation preserves valuable information regarding the development and reception of an original text (Tov, 2015). They can also be used for systematic exploration of syntactic structures between various types of related multi-lingual texts, including translations, redactions, and commentaries (Dorival, 2022).

In this paper we present a UD treebank of sections of the Septuagint produced by projecting and correcting the syntactic structure from the parallel text found in treebank presented in Swanson and Tyers (2022). Section 2 describes the texts used, Section 3 describes the annotation process and some specific syntactic considerations that came up in the process, Section 4 discusses quality metrics on the final treebank, Section 5 describes some preliminary investigations into future improvements of the methodology, and Section 6 concludes.

2. Corpus

Our base text for this work is the Codex Alexandrinus, one of the earliest more-or-less complete copies of the Septuagint. We obtained a morphologically annotated copy of the text from John Barach at GreekDoc.com[1]. That website was created as an educational resource and is structured such that each portion of the text is an HTML page and each word in the text is a link to a dictionary entry listing the headword, morphological features, and translations of the form in question. An example of the HTML representation is given in Figure 1.

Book	Sentences	Tokens	Words
Genesis	1,491	37,099	37,106
Ruth	85	2,400	2,403
Total	1,576	39,499	39,509

Table 1: Sizes of the texts in this treebank.

The Hebrew Scriptures are made up of 39 books, of which we have annotated the 2 which are currently available in the Ancient Hebrew treebank. The sizes of these two documents are presented in Table 1. They comprise roughly 7% of the total text of the Septuagint.

3. Annotation Process

After converting the HTML documents to CoNLL-U, we followed a process similar to that of Agić et al. (2016). We extracted the sequence of lemmas from each sentence in the Greek text and paired it with the sequence of lemmas from the corresponding sentence in the Ancient Hebrew treebank (skip-

[1] Now available at `https://greekdoc.github.io`.

LREC-COLING 2024, pages 13074–13078

```
<span class="num"><a id="v1">1</a></span>
<a href="../lexicon/en.html#en" title="in, on, by, with, to">Ἐν</a>
<a href="../lexicon/arc.html#arch6" title="beginning, ruler, office">ἀρχῇ</a>
<a href="../lexicon/epo.html#epoihsen" title="to do, make">ἐποίησεν</a>
<a href="../lexicon/o.html#o(" title="the; oh">ὁ</a>
<a href="../lexicon/qe.html#qeos" title="god">θεὸς</a>
<a href="../lexicon/to.html#ton" title="the">τὸν</a>
<a href="../lexicon/ou.html#ouranon" title="heaven, sky">οὐρανὸν</a>
<a href="../lexicon/kai.html#kai"
    title="and, also, even, then, next">καὶ</a>
<a href="../lexicon/th.html#thn" title="the">τὴν</a>
<a href="../lexicon/gh.html#ghn" title="earth">γῆν</a>.
```

Figure 1: The HTML representation of Genesis 1:1 Ἐν ἀρχῇ ἐποίησεν ὁ θεὸς τὸν οὐρανὸν καὶ τὴν γῆν. /en arxe epoiesen ho theos ton ouranon kai ten gen/ "In the beginning, God created the heavens and the earth."

ping punctuation, since there is little, if any, correspondence between the two systems). To these lemma sequences we applied the Eflomal word aligner (Östling and Tiedemann, 2016). Every arc in a Hebrew tree for which both the head and the dependent were aligned to Greek words was then copied to the Greek tree. This differs from Agić et al. (2016) in that their word aligner returned probabilities which they then used to place probabilities on the arcs that they produced, whereas we effectively tree the probability of every possible alignment as eith 100% or 0%. We opted for this binarization because none of the subsequent stages of the process support probabilities and because the primary purpose of projecting the trees was to save time for the annotators and not necessarily to achieve the maximum possible accuracy.

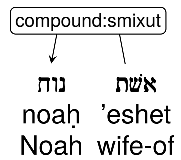
The projected trees were then uploaded to Arborator-Grew (Guibon et al., 2020), a dependency annotation platform which supports using queries to systematically rewrite various constructions. We used these queries to handle a variety of structural mismatches between Hebrew and Greek in conjunction with manual review and correction of all sentences by one of two editors.

After the manual correction, we compared the initial projected trees with the final versions and calculated the Cohen's Kappa (Cohen, 1960) as a measure of how accurate the process was. The result was 0.580 for head attachment and 0.503 for label selection, suggesting a usable, though somewhat limited baseline, which is in keeping with our experience of editing it. Unfortunately, the application of rewrite rules and manual editing are interleaved so as to prevent us from performing a similar calculation on the result of our systematic transformations.

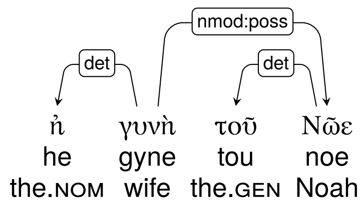
The following subsections discuss some of the specific syntactic constructions that came up in this process.

3.1. Compound vs Nominal Modifier

The Ancient Hebrew text makes frequent use of a construction called "smixut" which is analyzed in the treebank with a subtype of the `compound` relation. When translated into Greek, these frequently appear as possessive constructions, as shown in (1) and (2).

(1)


"Noah's wife" (Hebrew)

(2)


"Noah's wife" (Greek)

Both of these phrases mean "Noah's wife", but they express it with substantially different UD relations. We can handle this mismatch with a rule like (3).

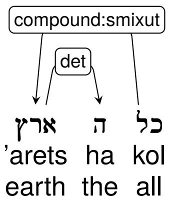
(3)

```
rule r1 {
  pattern {
    e: H -[compound:smixut]-> D;
    D[Case=Gen];
  }
  commands {
    e.label = "nmod:poss";
  }
}
```

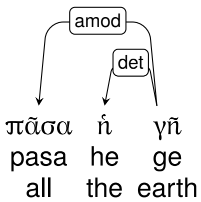
This finds an edge labeled `compound:smixut` where the dependent is in the genitive case and changes the label to `nmod:poss`.

3.2. Quantifiers

A frequent quantifier in the text is the Hebrew כל /kol/ “all”, which is typically rendered in Greek as πᾶς /pas/ “all”. While כל is a noun and typically appears in smixut constructions, πᾶς is an adjective and the appropriate relation is generally `amod`.

(4)


“the whole earth” (Hebrew)

(5)


“the whole earth” (Greek)

Both of these mean “the whole earth”, but the constructions have different headedness, which can be adjusted using a rule such as (6).

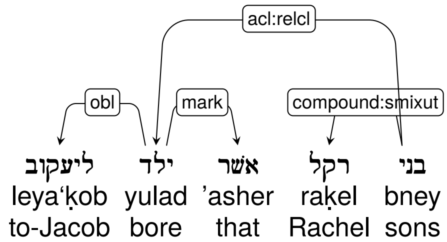
(6)

```
rule rl {
  pattern {
    P[lemma="πᾶς"];
    e: P -[compound:smixut]-> N;
  }
  commands {
    del_edge e;
    shift_in P ==> N;
    shift_out P ==> N;
    add_edge N -[amod]-> P;
  }
}
```

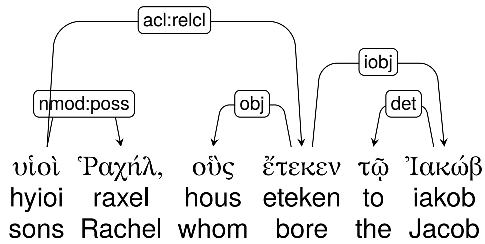
This locates any instance of the adjective πᾶς with a `compound:smixut` dependent (presumably a noun). It then deletes the `compound:smixut` edge and changes the parent of πᾶς to instead be the parent of the noun and changes any other dependents of πᾶς (such as a preposition) to be dependents of the noun as well. It then makes πᾶς a dependent of the noun with relation `amod`.

3.3. Relative Clauses

Relative clauses in the Hebrew text are frequently introduced with the subordinating conjunction אשר /’asher/ “that” and do not have distinct relative pronouns. In Greek, on the other hand, the relative clause typically begins with a relative pronoun which, due to aligning by lemmas, means that the predicted label for relative pronouns is usually `mark` when it should actually be a nominal argument such as `nsubj` or `obj`. An example is given in (7) and (8).

(7)


“the sons of Rachel, whom she bore to Jacob” (Hebrew)

(8)


“the sons of Rachel, whom she bore to Jacob” (Greek).

Here we need to find pronouns attached with `mark` and change that label to the correct argument relation, which we can largely accomplish based on the morphological case of the pronoun. Thus we will end up with several rules like (9).

(9)

```
rule rl {
  pattern {
    e: V -[mark]-> P;
    * -[acl:relcl]-> V;
    P[PronType=Rel,Case=Acc];
  }
  commands {
    e.label = "obj";
  }
}
```

This finds an accusative relative pronoun whose relation is `mark` and whose parent’s relation is `acl:relcl` and changes the relation from `mark` to `obj`.

Feature	Agreement
Heads	0.868
Relation Labels	0.813

Table 2: Inter-annotator agreement using Cohen's Kappa (Cohen, 1960).

4. Evaluation

Of the 54 chapters that comprise the two books in our corpus, 3 were corrected by both annotators (Ruth 2-4) and the inter-annotator agreement scores are presented in Table 2. The score of 0.868 for head selection indicates a fairly good agreement on structure and the score of 0.813 for labels suggests that it may be advisable to expand and clarify some of the Ancient Greek-specific annotation guidelines.

5. Future Work

Two potential avenues present themselves for improving on our methodology for future expansion of this treebank: trying to improve the word alignments and making the transformation rules reproducible.

In our current setup, each document is word-aligned using only the text found in the document itself. However, Eflomal supports saving the alignment probabilities from one run to be used for subsequent runs. Thus we can align the entirety of the Hebrew text with the Greek, even in the absence of full annotations, and in theory get more accurate results. The results of our initial attempt are listed in Table 3. This approach gave a small improvement in head attachment for Ruth but seems to otherwise have had a negligible impact.

Arborator-GRew does not provide a way to save transformation rules for future use, which is somewhat limiting when new texts are added to a corpus. However, GRew, the component which processes the rules, is also available as an offline system, allowing some of the transformations to be done as a preprocessing step. We assembled a set of 12 rules similar to the ones we ran during the annotation process and applied them to the original input files. The results are in Table 3. We found a moderate improvement on head attachment in Genesis ($+0.04\kappa$) and a larger one in Ruth ($+0.13\kappa$). Of the 12 rules, 7 only adjust labels without editing the tree structure, and thus the scores improve even more on relation labels, with $+0.14\kappa$ for Genesis and $+0.15\kappa$ for Ruth. More rules could be added, which would likely lead to even greater improvements.

	Genesis		Ruth	
	Head	Label	Head	Label
Original	0.581	0.504	0.633	0.540
Large	0.580	0.505	0.641	0.539
Rules	0.622	0.639	0.774	0.690

Table 3: The quality of the predicted annotations for the current starting point ("Original"), attempting to improve the alignments by adding more text ("Large"), and with a set of rules being applied prior to any manual intervention ("Rules"). All scores are Cohen's Kappa relative to the gold standard annotations.

6. Conclusion

In this paper we have presented our new Ancient Greek treebank. We discussed the process of creating it by alignment and projection from the parallel Ancient Hebrew text and described the semi-automated means of correcting those projections, which achieved acceptable levels of inter-annotator agreement. In addition, we have begun exploring some avenues that have the potential to substantially improve the quality of the projected trees and thus speed up future expansion of this treebank.

7. Acknowledgements

We would like to thank John Barach for giving us permission to use the underlying data. We would also like to thank Robert Pugh for helpful feedback on a draft of this paper.

References

Željko Agić, Anders Johannsen, Barbara Plank, Héctor Martínez Alonso, Natalie Schluter, and Anders Søgaard. 2016. Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics*, 4:301–312.

Kenneth A Cherney. 2014. *Allusion as translation problem: Portuguese versions of second Isaiah as test case*. Ph.D. thesis, Stellenbosch: Stellenbosch University.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Gilles Dorival. 2022. The septuagint between textual criticism and redactional criticism. In *The Hebrew Bible Manuscripts: A Millennium*, pages 175–188. Brill.

Hanne Eckhoff, Kristin Bech, Gerlof Bouma, Kristine Eide, Dag Haug, Odd Einar Haugen, and Marius Jøhndal. 2018. The proiel treebank family: a standard for early attestations of indo-european languages. *Language resources and evaluation*, 52:29–65.

Gaël Guibon, Marine Courtin, Kim Gerdes, and Bruno Guillaume. 2020. When collaborative treebank curation meets graph grammars. In *LREC 2020-12th Language Resources and Evaluation Conference*.

Jeremy G Kahn, Matthew Snover, and Mari Ostendorf. 2009. Expected dependency pair match: predicting translation quality with expected syntactic structure. *Machine Translation*, 23:169–179.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajic, Christopher D Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043.

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with Markov Chain Monte Carlo. *Prague Bulletin of Mathematical Linguistics*, 106:125–146.

Daniel Swanson and Francis Tyers. 2022. A universal dependencies treebank of ancient hebrew. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2353–2361.

Emanuel Tov. 2015. *The text-critical use of the Septuagint in biblical research*. Penn State Press.