Towards Named-Entity and Coreference Annotation of the Hebrew Bible

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

Named-entity annotation refers to the process of specifying what real-world (or, at least, external-to-the-text) entities various names and descriptions within a text refer to. Coreference annotation, meanwhile, specifies what context-dependent words or phrases, such as pronouns refer to. This paper describes an ongoing project to apply both of these to the Hebrew Bible, so far covering most of the book of Genesis, fully marking every person, place, object, and point in time which occurs in the text. The annotation process and possible future uses for the data are covered, along with the challenges involved in applying existing annotation guidelines to the Hebrew text.

Keywords: Ancient Hebrew, coreference, named-entity

1. Introduction

Coreference annotation is the process of marking whether or not two words or phrases in a document refer to the same document-external entity (whether real or imagined), which is very useful for information retrieval.

Named-entity and coreference annotation allows scholars, language instructors and students to view and search a version of the text that goes beyond lemmas to represent the real-world entities that tie the text together. Searches in this corpus can provide all mentions to a real-world entity, not only instances of a particular lemma, and facilitate linguistic inquiries at the syntax-semantics interface–where the entity type affects its usage in the sentence.

There are two main ways of accomplishing such annotations: links and clusters (Nedoluzhko et al., 2022). With linked annotations, each marked phrase is attached to another phrase (generally the nearest preceding one) with which it corefers. With clusters, on the other hand, a separate list of entities is created and each phrase is tagged as referring to one of those entities.

This paper presents the creation of a corpus of Ancient Hebrew annotated for co-reference using the cluster method, which thus also serves a set of named-entity annotations as well.

Ancient Hebrew is a Semitic language formerly spoken in the region that is now Israel and Palestine in the first and second millennia BC which survives to the present day in liturgical contexts.

The available corpus of texts in Ancient Hebrew (as distinct from Mishnaic Hebrew, a daughter language used by Jewish scholars during the Middle Ages) consists primarily of versions of the documents that now make up the Hebrew Bible. The standard versions of these texts contain 300-500 thousand words, depending on tokenization.

Lemmatization and part-of-speech tagging for the entirety of this corpus were completed by Peursen et al. (2015) and the first section of the corpus (30 thousand words) were syntactically annotated using the Universal Dependencies framework in Swanson and Tyers (2022). In this paper, we present the results of a pilot study on expanding the Universal Dependencies treebank to include co-reference and named-entity annotations.

The paper is organized as follows: Section 2 discusses the annotation scheme and the tools used in the annotation process. Section 3 provides a variety of statistics concerning the distribution of the resulting annotations. Section 4 describes the steps taken to measure the annotation quality and Section 5 concludes.

2. Annotation

In this project, we followed the CorefUD standard (Nedoluzhko et al., 2022), which is designed to be compatible with the Universal Dependencies file format. This meant that the annotations are done such that each phrase ("mention") points to an entry in a separate list rather than to a preceding (or, perhaps, following) phrase to with which it corefers.

CorefUD does not, however, provide definitions for what should and should not be included in the annotations. For this we used a subset of the co-reference guidelines used in the Universal Dependencies English GUM treebank (Zeldes, 2017), specifically the criteria for being a mention, the list of entity types, and the criteria for identifying two mentions as coreferential[1].

[1]The GUM guidelines can be found at https://

LT4HALA 2024@LREC-COLING-2024, pages 36–40
25 May, 2024. © 2024 ELRA Language Resources Association: CC BY-NC 4.0

Following the definitions in GUM, every noun phrase, proper noun, and pronoun in the corpus (including nested phrases) was included as a mention, apart from interrogative pronouns and a handful of a few language-specific constructions deemed to be non-referential, such as לְבַדּוֹ /levado/ "alone" (literally "to his separation"), where the central element בַּד /vad/ is a noun and thus forms a noun phrase with the possessive pronoun וֹ /o/, but the phrase has no meaningful referent. In this instance the pronoun is marked as coreferential with the appropriate entity (usually a person), while the noun is not part of any mention. Demonstrative adverbs such as שָׁם /sham/ "there" are also mentions, as are clauses and coordinated noun phrases which are referred back to. A consequence of this is that the resulting list of entities includes things that would not be found in any external ontology, such as the individual animals being sacrificed in a particular passage or an entity for a person's name separate from the entity for the person themself. The latter case occurs several times when describing the birth of a child, where text typically has some variant of "And they called his name 'Isaac'." Here *his* refers to Isaac, while *his name* and *'Isaac'* refer to Isaac's name rather than to Isaac himself.

Each entity is assigned one of the 10 types used in GUM and CorefUD. These are listed in Table 1. The definitions have been retained from GUM, but some names have been changed solely so that no two types have the same first letter, allowing us to use single letter mnemonics in our data files and annotation interface as described below.

The coreference guidelines from GUM which were used in this project primarily pertain to the circumstances under which copular predicates are or are not considered to corefer with their subjects.

To produce the annotations, the rule-based coreferencer Xrenner (Zeldes and Zhang, 2016) was applied to the treebank. The mentions it detected were exported, but our initial investigation found that the accuracy of its coreference labels was too low to be particularly helpful, so we opted to discard these. A simple terminal interface was then constructed in Python which displays a mention and its immediate context to the annotator who can then choose to label it with an existing entity or create a new entity. Entities can be referred to by ID, which consists of the first letter the entity type and a number counting up sequentially from the beginning of the corpus. Thus, when this project is expanded to include the entire UD treebank, the first three Person entities will be God (p1), Adam (p2), and Eve (p3). Many of the entities also have human-readable names, for which the annotation interface provides an autocomplete function (adding a name is optional if the annotator is confident that the entity in question is only referred to once). An example of the interface is given in Figure 1.

All the code used in this project is freely available and can be found with the data at https://github.com/mr-martian/hbo-UD. The data will also be converted to the CorefUD format and included in the upcoming version 1.2 release.

3. Corpus Statistics

The underlying corpus of the present project is a portion of the UD_Ancient_Hebrew-PTNK treebank (Swanson and Tyers, 2022) as of Universal Dependencies version 2.13 (Nivre et al., 2020), specifically containing the test and development sets and half of the training set. The size of this corpus is summarized in Table 2. This comprises the first 40 chapters of Genesis.

The coreference annotations label over 10,000 mentions referring to almost 1500 distinct entities. The distribution of entities and mentions by type is given in Table 3. The most common entity type is Person, which covers roughly 35% of the entities and 70% of the mentions. The least common, meanwhile, is Vegetation, at 1.5% of the entities and 0.6% of the mentions.

Eleven entities are referred to more than 100 times. All but one of them are Persons: The patriarchs Abraham (524), Isaac (208), Jacob (567), and Joseph (173), God (437), Jacob's brother Esau (189), Jacob's uncle Laban (156), Abraham's wife Sarah (122), Isaac's wife Rebecca (113), and one of Abraham's servants (102). The only location with more than 100 mentions is "the world" (149). Together these 11 entities total 2740 mentions, 37% of the total.

At the other end, there are 867 entities which are only mentioned once, which is 58% of all entities and 12% of all mentions.

4. Evaluation

One of the 40 chapters (specifically, Genesis 6) was chosen at random to be annotated twice. We measured agreement using the metrics provided by the corefUD scorer[2], which is an evaluation tool based on the Universal Anaphora Scorer (Yu et al., 2022), but adapted to the corefUD format. Each metric compares a reference document to a system output, so we ran the scorer with each annotator as the reference and averaged the resulting scores. The results are shown in Table 4. In addition, we give an analysis of the raw agreement rates on span selection, coreference, and entity type, since

wiki.gucorpling.org/gum/entities.

[2] https://github.com/ufal/corefud-scorer

GUM Label	Our Label	Examples
person	person	God, Abraham, the messenger of God
place	location	Bethel, Egypt, in the field
organization	nation	the Egyptians, the army of the Philistines
object	inanimate	a water-skin, a gold nose-ring
event	event	a feast, this thing that you have done
time	time	forever, the morning after the feast
substance	substance	the water of the well, the gold of that land
animal	creature	Abraham's donkeys, seven fat cows
plant	vegetation	the Tree of Life, a bush
abstract	abstract	his love for Rachel, favor in your sight

Table 1: The 10 entity types used in the corpus and how they relate to the GUM entity types. The names of the types used in the current corpus were chosen so as to be uniquely identifiable by their first letter.

```
Masoretic-Genesis-2:23-hbo
ויאמר האדם זאת הפעם עצם מעצמי ובשר מבשרי לזאת יקרא אשה כי מאיש לקחה זאת׃
עצם מן | ה פעם | אדם זאת
53:6-53:7 u1 (_)
> setnew t t:Adam-seeing-Eve
New ID: t122
```

Figure 1: The interface of the annotation tool. The first line gives the id of the sentence in the treebank. The second gives the full text of the sentence (in this case it reads "And the man said 'This one, now, is bone from my bone and flesh from my flesh. Because of this she shall be called "woman" because from man she was taken.'.") and the third gives the lemmas of each word in the current mention (here הפעם /hapa'am/ "now") along with the nearest two words on either side. The next line is the internal representation of the mention. `53:6-53:7` indicates that the mention begins at the 6th word of sentence 53 and ends at the 7th. `u1` is the current entity associated with this mention, in this case the first unknown and _ is the human-readable name of the entity, which is empty, since this is an unknown. `>` is a prompt for a command and the command here entered assigns this mention to a newly-created Time (`t`) entity with the name "t:Adam-seeing-Eve", which turns out to be the 122nd time entity created in this corpus.

	UD	CorefUD	Used
Sentences	1,579	1,161	73.5%
Words	39,036	28,485	73.0%
Tokens	26,846	19,621	73.1%

Table 2: Statistics about the UD_Ancient_Hebrew-PTNK treebank which was formed the basis of this project as of UDv2.13 and the resulting coreference corpus. The final column gives the proportion of the UD data which was used in the present work.

Entity type	Entity count	Mention count
Person	477	4842
Location	187	833
Abstract	218	429
Inanimate	173	372
Creature	100	276
Nation	73	259
Time	150	227
Substance	40	94
Vegetation	30	72
Event	47	69
Total:	1495	7473

Table 3: The frequency of entities and mentions in the corpus by entity type, sorted by number of mentions.

these 3 areas more directly show ways of improving the annotation process. A summary of these agreement rates is also given in Table 4.

4.1. Span Selection

The automated annotations consisted of 202 spans. Given the actions of 'annotate', 'delete', and 'modify', the two annotators agreed in 179 cases (88.61%). An analysis of the disagreements found that Xrenner overgenerates spans for entity mentions and the annotation guidelines were unclear on the proper treatment of some phenomena.

For example, Xrenner gives some determiners separate mentions due to part-of-speech tags. In (1), the word כל (kol "all, whole") is a noun, both

Measure	Agreement Rate	
Spans	179 / 202	88.61%
Spans (corrected)	188 / 202	93.07%
Coreference	129 / 147	87.76%
Entity Type	121 / 147	82.31%
LEA	70.02	± 1.15
MUC	81.44	± 1.04
B^3	73.55	± 0.69
CEAFe	62.66	± 1.11
CEAFm	77.73	± 0.87
BLANC	78.32	± 0.92
CoNLL	72.55	± 0.94

Table 4: Inter-annotator agreement statistics for Genesis chapter 6. "Spans" and "Spans (corrected)" refers to filtering of the original list of spans before and after an automated correction step was added (see Section 4.1). "Coreference" refers to the rate of agreement on which spans are and are not the same entity (Section 4.2). And "Entity Type" refers to whether the types of the entities match (Section4.3). The other scores are the F1 scores reported by the corefUD scorer. The scores are not symmetric with respect to which set of annotations is the reference, so we report the average (with variation) of the two directions.

etymologically and in the UD part-of-speech tags, and thus Xrenner creates mentions for both "the whole land" and "the land", when only the former should be annotated.

(1) ההוא הארץ כל
ה-הוא ה-ארץ כל
3SG.M-DEF land-DEF whole

"the whole of that land"

Similarly, ההוא (hahu' "the-him, that") is the 3rd person singular masculine pronoun with a definite article, a construction which serves as a demonstrative rather than as a referential pronoun. Thus, Xrenner produces a distinct mention for "that" in addition to "that land".

Fortunately, these issues, and a related one for numerals, can be fixed with an automated preprocessing step. Further, they can be automatically filtered from the Xrenner output, thus reducing annotator effort and risk of error.

Automatic correction took care of 9 disagreements, raising the agreement rate for span identification to 188 / 202 (93.07%).

4.2. Coreference

147 spans were given a label by both annotators. We calculate coreference agreement as follows: Given that annotator 1 applied a particular label to a set of spans, how many of those spans did annotator 2 label as coreferential? For example, if annotator 1 assigned a label of `i12` to 5 spans and annotator 2 assigned `s9` to 3 of the same spans and `c4` and `c5` to the other 2, we would calculate the coreference agreement by saying that annotator 2 agrees that 3 / 5 (60%) of spans are coreferential to one another (the particular labels being ignored for this measure).

Using the measure on the test sample, we observe an agreement rate of 129 / 147 (87.76%).

An example of an instance where the annotators disagreed was in Genesis 6:2, which refers בני האלהים /beney ha'elohim/ "the sons of God/the gods". Both annotators agreed on the coreference of the larger phrase as being a mysterious group not mentioned elsewhere, but one interpreted the nested mention as one of the names of God while the other read it as a plural noun referring to some other group of supernatural figures. The released version of the data takes the first interpretation, somewhat arbitrarily, pending a further analysis of evidence beyond the local lexical and syntactic context, since neither of those provide grounds for a decision.

4.3. Entity Types

Of the 147 spans annotated by both annotators, there were 26 cases where the entity type differed between them, giving an agreement rate of 121 / 147 (82.31%). The primary source of disagreement (14 of the 26 differences) was due to an unclear definition of the "nation" ("organization") entity type. It was sometimes used to refer to any group, though the intended use was for a group of people such that changing the specific members does not change the identity of the group (for example, the people of Egypt or the Philistine army). Thus, one annotator marked the set of all humans and animals as "nation" while the other marked it as "creature" (the released data has "creature"). Existing entities of this type have been reviewed and corrected as necessary.

5. Conclusion

In this paper we have presented a corpus of coreference annotations for Ancient Hebrew along with a description of the annotation guidelines and process, and distribution statistics distribution of various features in the text. We also presented the inter-annotator agreement of the text with discussion of methods to increase agreement via clarifications of the guidelines and improvements to the annotation pipeline.

In the future, we plan to expand the corpus to

cover the rest of the Hebrew Bible. In addition, there are several other types of annotations which commonly accompany co-reference, such as annotating relationships between entities (e.g. bridging, or part-whole relationships), which can be partially derived from our entity naming process, and linking the entity IDs to external sources, such as Wikipedia. Such extensions would greatly enhance the usefulness of this resource by enabling more complex querying of the data.

Acknowledgements

We would like to think Naomi Brokema and Amir Zeldes for discussing particularly tricky annotations decisions and thus helping to clarify the annotation scheme.

6. Bibliographical References

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

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, Amir Zeldes, and Daniel Zeman. 2022. CorefUD 1.0: Coreference meets Universal Dependencies. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4859–4872, Marseille, France. European Language Resources Association.

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.

W.T. van Peursen, C. Sikkel, and D. Roorda. 2015. Hebrew text database ETCBC4b.

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, Marseille, France. European Language Resources Association.

Juntao Yu, Sopan Khosla, Nafise Sadat Moosavi, Silviu Paun, Sameer Pradhan, and Massimo Poesio. 2022. The universal anaphora scorer. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4873–4883, Marseille, France. European Language Resources Association.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Amir Zeldes and Shuo Zhang. 2016. When annotation schemes change rules help: A configurable approach to coreference resolution beyond OntoNotes. In *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)*, pages 92–101, San Diego, California. Association for Computational Linguistics.