Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2353–2361
Marseille, 20-25 June 2022

A Universal Dependencies Treebank of Ancient Hebrew

Daniel G. Swanson, Francis M. Tyers
Department of Linguistics,
Indiana University,
{dangswan,ftyers}@iu.edu

Abstract

In this paper we present the initial construction of a Universal Dependencies treebank with morphological annotations of Ancient Hebrew containing portions of the Hebrew Scriptures (1579 sentences, 27K tokens) for use in comparative study with ancient translations and for analysis of the development of Hebrew syntax. We construct this treebank by applying a rule-based parser (300 rules) to an existing morphologically-annotated corpus with minimal constituency structure and manually verifying the output and present the results of this semi-automated annotation process and some of the annotation decisions made in the process of applying the UD guidelines to a new language.

Keywords: treebank, UD, Hebrew

1. Introduction

The Hebrew Scriptures are a collection of 39 books primarily written in the first millennium BC in Ancient Hebrew (with a few passages in Aramaic) which were arranged and codified in their current form over the course of the first millennium AD. They are also known as the Tanakh, an acronym of the Hebrew names of the 3 main divisions: תורה /torah/ "law"[1], נבאים /nevi'im/ "prophets" (a category which also includes several books of narrative history), and כתבים /ketuvim/ "writings".

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.

In this paper we present a UD treebank containing the books of Genesis and Ruth with the intent that, like other ancient language treebanks such as PROIEL (Haug and Jøhndal, 2008; Eckhoff et al., 2018), it can serve as a resource for studying both the Tanakh as a document and also the development of Hebrew syntax through the centuries. In addition, we chose to begin with these books in part because they can be compared with the Peshitta, a Coptic translation of the Bible which is partially available in UD (Zeldes and Abrams, 2018).

Section 2 describes the text and morphological annotations used, Section 3 presents how the treebank was created, Section 4 discusses some of the more challenging constructions to annotate, Section 5 provides statistics about treebank produced, and Section 6 concludes.

2. The Hebrew Corpus

The Biblia Hebraica Stuttgartensia Amstelodamensis (BHSA) is a complete copy of the text of the Tanakh with morphological annotations which is maintained by the Eep Talstra Centre for Bible and Computer (Peursen et al., 2015).

The corpus is stored as a table, where each row is a syntactic node, whether a word, phrase, clause, or sentence. The columns of the table specify various features of these these nodes such, for words, lemma, part of speech, person, gender, number, and whether there is a space before the following word. There are also columns specifying which larger nodes the node represented by a particular row is a part of. An example of this structure can be found in Table 1.

Phrases are contained in clauses and clauses in sentences. Phrases and clauses are not necessarily contiguous, though sentences are. An example of the structure when they are not contiguous is given in Table 2. This structure can be used to construct a rudimentary constituency tree, but with a very flat structure. See Figure 1 for an example of such a tree.

3. Annotation Process

Our annotation process consisted of the creation of an automated procedure for adjusting and converting the BHSA tokenization and part-of-speech tags to match the UD annotation guidelines and then passing this through a rule-based parser and manually validating the output.

3.1. Tokenization

The BHSA splits words on spaces except for a handful of proper nouns. It additionally separates the conjunction ו /ve, va, u/ "and", prepositional prefixes, and the definite article ה /ha/. However, it does not split off pronominal suffixes[2]. Thus an example of the maximal

[1] Instances of Hebrew script in this paper are followed by a transliteration in slashes according to the ALA-LC scheme (Barry, 1997) and an English translation in quotes.

[2] Pronominal suffixes in Hebrew can attach to prepositions as in לו /lo/ "to him", to nouns as possessors as in יד /yad/ "hand" vs ידו /yado/ "his hand", or to verbs as direct object as in לשׁהמר /lishmor/ "to guard" vs לשׁמרו /lishmero/ "to guard him".

id	type	word	space	POS	number	person	function	phrase	clause	sent
1	word	ב /be/	no	prep				12	16	17
2	word	רֵאשִׁית /reshit/	yes	subs	sg			12	16	17
3	word	בָּרָא /bara'/	yes	verb	sg	p3		13	16	17
4	word	אֱלֹהִים /'elohim/	yes	subs	pl			14	16	17
5	word	אֵת /'et/	yes	prep				15	16	17
6	word	ה /ha/	no	art				15	16	17
7	word	שָׁמַיִם /shamayim/	yes	subs	pl			15	16	17
8	word	ו /ve/	no	conj				15	16	17
9	word	אֵת /'et/	yes	prep				15	16	17
10	word	ה /ha/	no	art				15	16	17
11	word	אֶרֶץ /'arets/	yes	subs	sg			15	16	17
12	phrase						Time		16	17
13	phrase						Pred		16	17
14	phrase						Subj		16	17
15	phrase						Objc		16	17
16	clause									17
17	sent									

Table 1: A segment of the data table of the BHSA showing Genesis 1:1. Each row of the table represents a linguistic unit (word, phrase, clause, or sentence), with the rows being sorted first from smallest to largest and then by order of occurrence within the text. The columns represent various features, with an empty cell indicating that that feature is not applicable to that node. The last three features indicate which rows are nodes which contain the current node.

words	word	וכל האדם	על־הארץ	הרמשׂ	כל־בשׂר	יגוע
	gloss	and all humans	upon the earth	that crawled	all flesh	perished
phrases	id	2	4	3	2	1
	type	NP	PP	VP	NP	VP
	function	Subject	Location	Predicate	Subject	Predicate
clauses	id	5	6	6	5	5
	relation		Attributive	Attributive		
sentences	id	7	7	7	7	7

Table 2: An example of the structure of the phrase, clause, and sentence annotations in the BHSA. The sentence is part of Genesis 7:21 יגוע כל־בשׂר הרמשׂ על־הארץ וכל האדם /yigẏa' kol-baśar haromeś 'al-ha'arets vekhol ha'adam/ "All flesh that crawled upon the earth and all humans perished." Here the relative clause "that crawled upon the earth" intervenes between two members of a list, causing both phrase 2 and clause 6 to be non-contiguous. Note also that phrase 2 "all flesh and all humans", phrase 3 "that crawled", and phrase 4 "upon the earth" are all entirely separate from one another and there is no hierarchical relation between them.

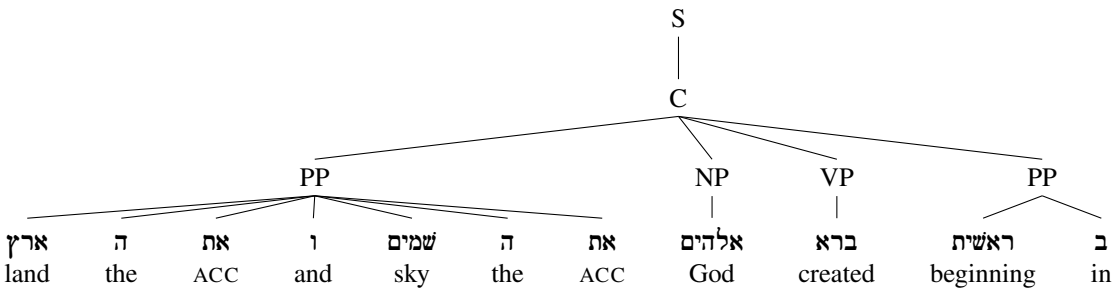


Figure 1: An example of the rather limited syntax encoded directly in the BHSA. Note especially the complete lack of internal structure in the second prepositional phrase. This is Genesis 1:1 בראשׁית ברא אלהים את השׁמים ואת הארץ /bereshit bara' 'elohim 'et hashamayim ve'et ha'arets/ "In the beginning, God created the heavens and the earth."

amount of splitting is (1)[3].

(1) **בבית ובידו**

ב	**ה**	**בית**	**ו**	**ב**	**ידו**
b	a[ha]	bayit	u	be	yad=o
in	DEF	house	and	in	hand=3SG.M

"in the house and in his hand"

Note that the definite article has been reduced to a change in the vowel of the preceding preposition and thus is visible in the text with vowels but not in the plain consonantal text.
In this work, however, we follow the Modern Hebrew UD treebank (Tsarfaty, 2013; McDonald et al., 2013) in also splitting off prepositional suffixes as in (2).

(2) **בבית ובידו**

ב	**ה**	**בית**	**ו**	**ב**	**יד**	**הוא**
b	a[ha]	bayit	u	be	yad	o[hu']
in	DEF	house	and	in	hand	3SG.M

"in the house and in his hand"

As shown in (2), we give the pronominal suffix the lemma of the corresponding independent pronoun. Additionally, in accordance with UD guidelines, we convert punctuation from a property of the preceding word to a full token so that it can be attached to the final dependency tree.

3.2. Part-of-Speech Tags

The conversion from BHSA POS tags to the POS tags used in UD is summarized in Table 3. The cases in which this conversion is not one-to-one are as follows:

3.2.1. Adjectives

Words tagged as adjectives in the BHSA are retagged as nouns if they have a pronominal possessor or participate in nominal compounding (see Section 4.1), such as in (3).

(3) **ויקח עשׂרה אנשׁים מזקני העיר**

ו	**יקח**		**עשׂרה**	**אנשׁים**
ṿe	yiḳaḥ		ʻesrah	ʼanashim
CCONJ	VERB		NUM	NOUN
and	3SG.M-take.IMPF		ten	man-PL
מ	**זקני**	**ה**	**עיר**	
mi	ziḳne	ha	ʻir	
ADP	ADJ/**NOUN**	DET	NOUN	
from	old-PL.CNST	DEF	city	

"He took 10 men from among the elders of the city." (Ruth 4:2)

[3]Although Hebrew text is written right-to-left, glossed examples in this paper present the words from left to right for readability.

BHSA	**Description**	**UD**
adjv	adjective	ADJ, NOUN
advb	adverb	ADV
art	article	DET, SCONJ
conj	conjunction	CCONJ, SCONJ
inrg	interrogative particle	ADV, PART
intj	interjection	INTJ
nega	negative particle	ADV
nmpr	proper noun	PROPN
prde	demonstrative pronoun	PRON
prep	preposition	ADP
prin	interrogative pronoun	PRON
prps	personal pronoun	PRON
subs	noun	NOUN, ADP, ADV, VERB
verb	verb	VERB, AUX, NOUN
prn*	pronominal suffix	PRON
punct*	punctuation	PUNCT

Table 3: Mapping of POS tags from BHSA to UD. The tags prn and punct are not actually present in the BHSA, but are inserted by the tokenization procedure described in Section 3.1

Here **זקני**, a form the adjective **זקן** /zaḳen/ "old", has the plural form of the nominal compound suffix, indicating that the following noun depends on it. As a result, we tag it as a noun.
Words denoting nationalities so frequently stand on their own rather than modifying a noun that the decision was made to also treat them as nouns in their own right, such as in (4).

(4) **בני יון כתים ודדנים**

בני	**יון**	**כתים**	**ו**
bne	yaṿan	kitim	ṿe
NOUN	PROPN	ADJ/**NOUN**	CCONJ
son-PL.CNST	Yavan	Kittite-PL	and
דדנים			
dodanim			
ADJ/**NOUN**			
Dodanite-PL			

"The sons of Yavan were the Kittites and the Dodanites." (Genesis 10:4)

3.2.2. Articles

The definite article is tagged as a subordinating conjunction when it attaches to a non-nominalized participle, such as in (5).

(5) וכל־הרמשׂ הרמשׂ על־הארץ

ו	כל	ה	רמשׂ	ה
ṿe	khol	ha	remeś	ha
CCONJ	NOUN	DET	NOUN	DET/**SCONJ**
and	all	DEF	creeper	DEF

רמשׂ	על	ה	ארץ
romeś	‘al	ha	’arets
VERB	ADP	DET	NOUN
creep.PART	upon	DEF	earth

“and every creeping thing that creeps upon the earth” (Genesis 7:14)

This situation is discussed in more detail in Section 4.1.

3.2.3. Conjunctions

Unlike UD, the BHSA does not make a distinction between coordinating and subordinating conjunctions. We treat ו /ve/ “and” and או /’o/ “or” as coordinating conjunctions. Other conjunctions such as the relative clause marker אשׁר /’asher/, כי /ki/ “that, because”, and אם /’im/ “if” are all tagged as subordinating.

3.2.4. Interrogative Particle

The BHSA category of interrogative particle covers both question words, which are tagged as adverbs (ADV), and the question marker ה /ha/, which is tagged as a particle (PART).

3.2.5. Nouns

Words that the BHSA treats as nouns are retagged in a variety of situations because the BHSA tagging is based more on etymology than on the current behavior of a word. As a result, some words such as אחר /’aḥer/ “after” and מאד /me’od/ “very” are retagged as prepositions and adverbs, respectively, such as in (6).

(6) לא־יודע אחרי־כן כי־כבד הוא מאד

לא	יודע		אחרי
lo’	yiṿada‘		’aḥare
ADV	VERB		NOUN/**ADP**
NEG	3SG-remember.IMPF.PASS		after

כן	כי	כבד	הוא	מאד
khen	ki	khavod	hu’	me’od
ADV	SCONJ	ADJ	PRON	NOUN/**ADV**
this	because	heavy	3SG.M	very

“[The abundance] will not be remembered after this because it [the famine] will be very great.” (Genesis 41:31)

In addition, there are two existential verbs, יש /yesh/ “there exists” and אין /’en/ “there does not exist”, which we tag as verbs (VERB), such as in (7).

(7) ולחם אין בכל־הארץ

ו	לחם	אין	ב	כל
ṿ	leḥem	’en	be	khol
CCONJ	NOUN	NOUN/**VERB**	ADP	NOUN
and	bread	NEG	in	all

ה	ארץ
ha	’arets
DET	NOUN
the	land

“And there was no bread in all the land.” (Genesis 47:13)

3.2.6. Verbs

Words tagged as verbs in the BHSA are also tagged as verbs in UD, except for the copula היה /hayah/, which is tagged as an auxiliary. Verbs in participial form are also sometimes treated as nominalized and so tagged as nouns. This case is described in more detail in Section 4.1.

3.3. Parsing

To produce dependency annotations, we constructed a rule-based parser using the VISL Constraint Grammar formalism (Bick and Didriksen, 2015) as it has been successfully used for prior annotation projects (Bick, 2005; Antonsen et al., 2010; Tyers and Sheyanova, 2017) and can easily process arbitrarily many annotation labels. Most of the annotation layers of the BHSA (excluding text-formatting directives) are converted to the Constraint Grammar input format. Tokens marking boundaries between phrases, clauses, and sentences are also inserted.

For this treebank, we have chosen to follow the traditional verse boundaries rather than the BHSA sentence boundaries since BHSA puts quotations and subordinate clauses expressing conditions or causes in separate sentences from their parent clauses. BHSA thus often annotates multiple sentences for a single verse, though the verse boundaries usually align with sentence boundaries. When they do not align, we automatically merge adjacent verses into a single tree.

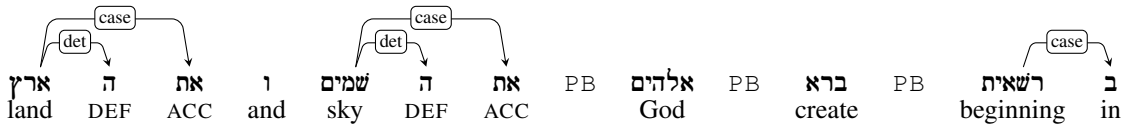
After splitting into sentences, a further pre-processing step uses the BHSA phrase function labels and clause ids to mark the phrase in each clause which most likely contains the root of that clause.

Each tree is then passed through a parser consisting of 307 Constraint Grammar rules. These include 121 head-assignment rules, 114 relation-assignment rules, and 63 rules manipulating the tags. Examples of these types of rules are shown in Table 4 and the process for an entire sentence is shown in Figure 2. The remaining 9 rules deal with instances where the BHSA tokenization disagrees with the UD guidelines and with removing the phrase boundaries before the output is converted to CoNLL-U format.

Finally, a script converts the dependency tree to CoNLL-U format together with all morphological an-

```
MAP @case Pr ;
SET AfterPrep = Det OR @case OR @nummod ;
SETPARENT @case TO (1* Noun OR PRON BARRIER (*) - AfterPrep) ;
MAP @det Det ;
SETPARENT Det TO (1 Noun OR ADJ OR PRON) ;
```

Attach prepositions to following nouns or pronouns as `case`, skipping any intervening determiners, prepositions, or numbers and attach determiners to immediately following nouns, adjectives, or pronouns with `det`.




```
MAP @cc (conj) - (Rela) ;
SETPARENT @cc TO (1* NPHead OR PPHead BARRIER PB) ;
```

Attach non-relative conjunctions to a following `NPHead` or `PPHead` (any noun not already attached to another noun) in the same phrase.

```
MAP @conj NPHead + HasConj + (/^\(ph\\d+\)$/r)
    IF (-1* NPHead + (VSTR:$1)) ;
SETPARENT NPHead + @conj + (/^\(ph\\d+\)$/r)
    TO (-1* NPHead - @appos - @conj + (VSTR:$1)) ;
```

In an `NPHead` has coordinating conjunction dependent and there is a preceding `NPHead` with the same phrase id, attach it to that as `conj`.


```
SETPARENT (/^\(c\\d+\)$/r) - CR TO (0* (VSTR:$1) + CR) ;
```

Attach the head of each phrase to the phrase with the same clause id which has the highest-precedence function label. In this case, the phrases are (from right to left) `Time` (temporal oblique), `Pred` (predicative verb), `Subj` (subject), and `Objc` (object). Since the verb is not a copula, it was marked as the clause root (`CR`) in a pre-processing step.

```
MAP @nsubj (Subj) - CR IF (p CR LINK NEGATE c @nsubj) ;
MAP @obj (Objc) - CR IF (p CR LINK NEGATE c @obj) ;
LIST OblIsh = Time Loca Modi Adju ;
MAP @obl OblIsh - CR IF (p CR) ;
```

Having attached various phrases within the clause, assign relations based on their function labels: `Subj` is `nsubj`, `Objc` is `obj`, and most other noun phrase functions are `obl`. The subject and object rules also check that there is at most one of each and other rules are applied if this is not the case.

```
MAP @root CR IF (NOT -1* CR - @advcl - @acl) ;
SETPARENT @root TO (@0 (*)) ;
```

Make a clause root the root of the sentence if there is no full clause preceding it.

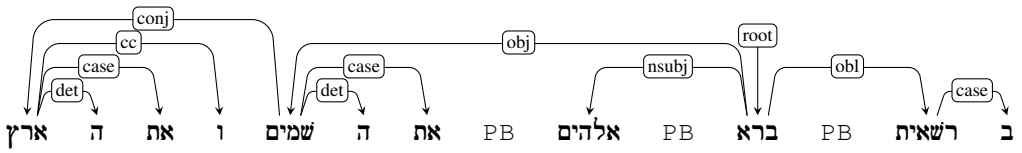


Figure 2: The process of parsing Genesis 1:1 **בראשית ברא אלהים את השמים ואת הארץ** /bereshit bara' 'elohim 'et hashamayim ve'et ha'arets/ "In the beginning, God created the heavens and the earth." `PB` represents a boundary between phrases. The corresponding constituency tree is given in Figure 1.

Rule Type	Count	Example
SETPARENT	121	`SETPARENT @cc (NOT p (*)) TO (1 (subs @conj)) ;` Set the head of a word with relation `cc` that does not already have a head to the immediately following word if that word has the part-of-speech tag `subs` (BHSA tag for common nouns) and the relation `conj`.
MAP	114	`MAP @obj (prn) IF (-1 (verb)) ;` Set the relation of a pronominal suffix to `obj` if the preceding word is a verb.
ADD	26	`ADD HasConj NPHead IF (NOT 0 HasConj) (c @cc) ;` If a word is marked as the head of a noun phrase (label `NPHead`) and has a dependent which is a coordinating conjunction (relation `cc`), then add the label `HasConj`.
SUBSTITUTE	37	`SUBSTITUTE (art) (conj retag:art) (CP Rela) ;` If a word is tagged as a determiner in a conjunction phrase (`CP`) which is functioning as a relativizer (`Rela`), then change its part-of-speech tag from `art` to `conj` (conjunction).

Table 4: Examples of the main types of Constraint Grammar rules used in the parser. SETPARENT creates a dependency arc, MAP assigns a relation, ADD assigns helper labels, and SUBSTITUTE changes annotations decisions in the BHSA corpus.

notations which correspond to features described either in the guidelines of the UD project in general or of the Modern Hebrew treebank in particular.

Every time the rules are changed, all sentences are reparsed and compared against the previously verified versions to ensure that there have been no regressions. All the code involved in this process is available on Github[4] under an open-source license.

4. Annotation Decisions

The most significant annotation challenges we encountered involved participles, a marker for direct quotations, and an emphatic construction using infinitives.

4.1. Participial Relative Clauses

Like other Semitic languages, Ancient Hebrew nouns have a form known as 'construct state' which is used when combining them with other nouns. For example, compare (8) and (9).

(8) יֵשׁ בנים לישׂראל

יֵשׁ	בנים	לישׂראל
yesh	ban-im	le-yiśra'el
exist	son-PL	to-Israel

"Israel had sons."

(9) בני ישׂראל

בני	ישׂראל
bn-e	yiśra'el
son-PL.CNST	Israel

"the sons of Israel"

In (9), the noun in construct state immediately precedes another noun, which determines its definiteness. Since only a definite article can be placed between two nouns in this construction, we follow the Modern Hebrew treebank (Tsarfaty, 2013; McDonald et al., 2013) in annotating this as a `compound` relation, with the first noun as the head. The equivalent construction in some other Semitic languages such as Akkadian (Luukko et al., 2020) has a strictly genitive function and thus uses `nmod:poss`. We do not take the latter approach because the Hebrew construction is more general than possession but there are no morphosyntactic criteria that would distinguish possessive instances from non-possessive ones.

However, a problem then arises with participles, which can appear as either piece of a construct phrase, as in (10) or with an argument structure comparable to that of finite verbs as in (11).

(10) הוא היה אבי ישׁב אהל

הוא	היה	אבי
hu'	hayah	'avi
3SG.M	be.PERF.P3.SG.M	father.CNST

ישׁב	אהל
yoshev	'ohel
dwell.PART.CNST	tent

"He was the father of those who live in tents." (Genesis 4:20)

[4] https://github.com/mr-martian/hbo-UD

(11) הדד בן־בדד המכה את־מדין בשׂדה מואב

הדד	בן	בדד	המכה
hadad	ben	badad	ha-makeh
Hadad	son.CNST	Badad	DEF-strike.PART

את	מדין	בשׂדה	מואב
’et	midyan	be-śadeh	moav
DEF.ACC	Midian	in-field.CNST	Moab

“Hadad son of Badad, the one who struck Midian in the fields of Moab.” (Genesis 36:35)

In (10), ישׁב /yoshev/ “dwell, inhabit” is a participle in a nominal compound construction, while in (11), המכה /hamakeh/ “strike” is followed by the definite direct object marker את /’et/, a definitely verbal construction, though it does also have a definite article, suggesting a nominal interpretation.

We concluded that the approach most consistent with the UD guidelines was to treat participles as nominalized if they occur in construct state and have no verbal argument structure since in such cases the morphology and the syntax are both nominal. Conveniently, the BHSA marks such participles as part of the surrounding noun phrase rather than as a verb phrase in a separate clause. Thus we tag participles in the same phrase as NOUN and attach them with compound while other participles are tagged as VERB and usually attached with acl. In the latter case, if the participle has a definite article, this is retagged as a subordinating conjunction (SCONJ) and attached with the usual relation of mark.

4.2. Quotations

Direct quotations in Biblical Hebrew are frequently preceded by לאמר /le’mor/, which is both etymologically and in the BHSA an infinitive of “say” with the prepositional prefix ל /le/, which occurs in many constructions involving infinitives in addition to marking the dative on nouns and pronouns.

If the לאמר is taken as a verb, then the question arises of whether the quotation should depend on the infinitival speaking verb immediately before it or on the finite one earlier in the clause. There is also the question of how לאמר should relate to the finite verb. Other verbs in this form are usually either controlled clausal complements (xcomp) or purpose clauses (advcl).

On the other hand, לאמר is not required in the sentence and when it is absent, the quotation has to depend on the finite verb. For consistency with this case, the quotation could always be attached to the finite verb and לאמר could be attached to the quote if it is present. These two alternative trees are shown in Figure 3.

In the end we decided to follow the lead of the Coptic Scriptorium treebank (Zeldes and Abrams, 2018) and analyze לאמר as a subordinating conjunction which depends on the following quotation.

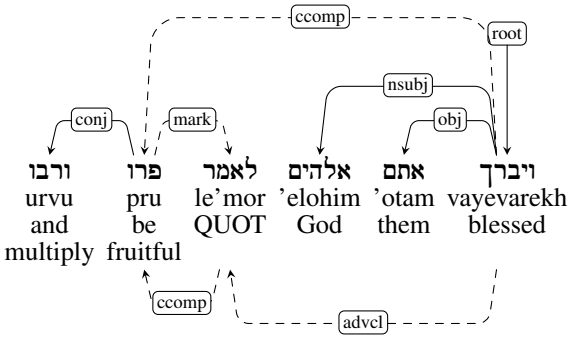


Figure 3: Two potential approaches according to the UD guidelines to analyzing the quotation marker לאמר in the sentence ויברך אתם אלהים לאמר פרו ורבו “God blessed them saying ‘Be fruitful and multiply!’.” (Genesis 1:22). The relation parataxis would be another possibility, if the verb for “bless” were not analyzed as a verb which introduces a quotation. (Internal structure of multi-word tokens is not shown.)

4.3. Infinitive Absolute

Biblical Hebrew has two verbal forms which are traditionally called infinitives, the infinitive construct and the infinitive absolute (named on analogy to the construct and absolute states of nouns discussed in Section 4.1). The infinitive construct is used in a variety of dependent clause constructions while the less common infinitive absolute occurs primarily in a single emphatic construction.

The typical appearance of the infinitive absolute is immediately preceding a conjugated finite version of the same verb, such as in (12).

(12) עשׂר אעשׂרנו לך

עשׂר	אעשׂרנו	לך
‘aśer	’a-‘aśr=enu	l-akh
tithe.INF	1SG-tithe.IMPF=3SG.M	to-2SG.M

“I will give a tenth to you.” (Genesis 28:22)

Here the infinitive absolute of the verb עשׂר /‘a,sar/ “tithe, give a tenth” emphasizes the immediately following conjugated form אעשׂרנו /’a‘aśrenu/ “I will give a tenth of it”.

The equivalent construction in Arabic takes accusative marking, suggesting that this should be analyzed as a nominal form, similarly to what happens in (13).

(13) הבה נלבנה לבנים

הבה	נלבנה	לבנים
havah	ni-lbenah	leven-im
JUSS	1PL-make.bricks.IMPF	brick-PL

“Let us make bricks.” (Genesis 11:3)

Here the verb נלבנה /nilbenah/ “we will make bricks” takes as a direct object a noun derived from the same consonantal root לבנים /levenim/ “bricks”.

Book	Trees	Words	Tokens
Genesis	1,494	36,741	25,282
Ruth	85	2,294	1564
Total:	1,579	39,035	26,846

Table 5: Size of the texts included in the Ancient Hebrew treebank.

Feature	Values	Occurrences
Aspect	2	1,965
Gender	2	16,621
HebBinyan	6	5,301
Mood	2	4,532
Number	3	20,018
NumType	1	477
Person	3	8,526
Polarity	1	277
PronType	3	4,300
Tense	1	2,244
VerbForm	3	5,347
Voice	1	49

Table 6: The 12 morphological feature categories included in the Ancient Hebrew treebank along with how many distinct values they have and how many words they appear on.

However, this causes problems with transitive verbs since the UD guidelines disallow having multiple words marked as objects (`obj`) of the same verb. Another option would be to mark these as `iobj` (indirect object), though the guidelines for `iobj` are mainly focused on recipients. In addition, the fact that this construction is entirely optional argues against using the core argument relations, which then suggests making them obliques (`obl`) instead.

We posed this question to the broader UD project and as of this writing that discussion has yet to reach a consensus.[5] As a temporary solution, we have attached the infinitive absolute to the following verb with `advmod` and correspondingly tagged them as `ADV`, which will be easy to update once the appropriate UD standard has been established.

5. Treebank Statistics

For this study we have parsed the books of Genesis and Ruth. Statistics about the texts can be found in Table 5. In addition, nouns, verbs, adjectives, and pronouns all have morphological annotations directly converted from the underlying corpus, the distribution of which is summarized in Table 6.

Validating the output of the conversion process was done entirely by the first author, making it impossible to report inter-annotator agreement. As an alternative measure of annotation consistency, we trained a parser using UDPipe (Straka et al., 2016; Straka and Straková, 2017) on the entire corpus and then parsed the corpus with that model. If the annotations are fully consistent, we would expect the model to perfectly reproduce the training data. The results are shown in Table 7.

[5]The discussion can be found at `https://github.com/UniversalDependencies/docs/issues/832`.

Metric	Score
UAS	96.27
LAS	94.65

Table 7: Results of training a parser to determine whether the annotations were consistent enough to be memorized by a model.

Error Type	Occurrences
Attachment	830
Verb Argument Identification	424
NP vs Clausal Modifier	159
Clause Type	144
Ellipsis	75
NP Structure	57
Other	399

Table 8: Occurrences of error types in the parser output.

The results are slightly lower than might be expected given the setup, so we also did an analysis of the most common errors in the parser output and found that the majority of them are due to distinctions that are not readily apparent in the information available to the parser. These include identifying whether a given prepositional phrase is a nominal modifier (`nmod`) of the immediately preceding noun or an oblique (`obl`) of the verb before that or determining where a quotation ends and the narrative resumes. Statistics about the identifiable classes of errors is shown in Table 8.

6. Concluding Remarks

In this paper we presented our approach to semi-automatically annotating an Ancient Hebrew treebank and discussed some of the difficulties involved in applying the UD guidelines to a new language.

The rule-based parser developed in this paper has been successfully applied to over 1500 sentences containing about 39000 words. It is thus likely that it can be applied with minimal adjustments to the remaining books of the Hebrew Scriptures, especially those from a similar time period and in a similar genre (narrative).

The treebank will be released as Ancient Hebrew-PTNK in UD version 2.10.

7. Acknowledgements

Our thanks to Amir Zeldes for valuable discussions of Hebrew syntax. Thank you also to Bryce Bussert, Matthew Fort, and Sandra Kübler for reading drafts of this paper.

8. References

Antonsen, L., Trosterud, T., and Wiechetek, L. (2010). Reusing grammatical resources for new languages. In *LREC*.

Barry, R. K. (1997). ALA-LC romanization tables-transliteration schemes for non-roman scripts. In *Library of Congress, 1997*.

Bick, E. and Didriksen, T. (2015). CG-3—beyond classical constraint grammar. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 31–39.

Bick, E. (2005). Turning constraint grammar data into running dependency treebanks. In *Civit, Montserrat & Kübler, Sandra & Martí, Ma. Antònia (red.), Proceedings of TLT 2005 (4th Workshop on Treebanks and Linguistic Theory, Barcelona*, pages 19–27.

Eckhoff, H., Bech, K., Bouma, G., Eide, K., Haug, D., Haugen, O. E., and Jøhndal, M. (2018). The proiel treebank family: a standard for early attestations of indo-european languages. *Language Resources and Evaluation*, 52(1):29–65.

Haug, D. T. and Jøhndal, M. (2008). Creating a parallel treebank of the old indo-european bible translations. In *Proceedings of the second workshop on language technology for cultural heritage data (LaTeCH 2008)*, pages 27–34.

Luukko, M., Sahala, A., Hardwick, S., and Lindén, K. (2020). Akkadian treebank for early neo-assyrian royal inscriptions. In *Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories*, pages 124–134, Düsseldorf, Germany, October. Association for Computational Linguistics.

McDonald, R. T., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K. B., Petrov, S., Zhang, H., Täckström, O., et al. (2013). Universal dependency annotation for multilingual parsing. In *Proc. of ACL*.

Nivre, J., de Marneffe, M.-C., Ginter, F., Hajic, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., and Zeman, D. (2020). Universal dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043.

Peursen, W. v., Sikkel, C., and Roorda, D. (2015). Hebrew text database ETCBC4b.

Straka, M. and Straková, J. (2017). Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August. Association for Computational Linguistics.

Straka, M., Hajič, J., and Straková, J. (2016). UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Tsarfaty, R. (2013). A unified morpho-syntactic scheme of stanford dependencies. In *Proc. of ACL*.

Tyers, F. and Sheyanova, M. (2017). Annotation schemes in North Sámi dependency parsing. In *Proceedings of the Third Workshop on Computational Linguistics for Uralic Languages*, pages 66–75.

Zeldes, A. and Abrams, M. (2018). The Coptic Universal Dependency Treebank. In *Proceedings of the Universal Dependencies Workshop 2018*, pages 192–201, Brussels.