A New Hebrew Universal Dependency Treebank: The First Treebank of Post-Rabbinic Historical Hebrew

Rachel Tal,[*,1,2] Shlomit Fuchs,[*,1] Orly Albeck,[2] Elisheva Brauner,[1,2]
Yitzchak Lindenbaum,[1,2] Ephraim Meiri,[1,2] Avi Shmidman[1,2]
[*]Equal contribution
[1]Bar-Ilan University, Ramat Gan, Israel
[2]DICTA, Jerusalem, Israel
rachel.tal@mail.huji.ac.il avi.shmidman@biu.ac.il
{shumital, orlyalbeck, efbrauner, yitzilindenbaum, ephraimmeiri}@gmail.com

Abstract

The corpus of post-Rabbinic historical Hebrew is a foundational corpus of Jewish heritage, containing over a billion words of legal, hermeneutical, and philosophic texts (and more). However, because the linguistic norms of the corpus diverge so often from that of modern Hebrew, the corpus cannot be computationally analyzed with existing Hebrew parsers. In order to fill this lacuna, we present the first Universal Dependencies corpus of post-Rabbinic historical Hebrew. The corpus comprises over 11,800 words, and we are pleased to release it to the community.

1 Introduction

The post-Rabbinic historical Hebrew corpus is comprised of over a billion words, authored across over a thousand years between the tenth and nineteenth centuries,[1] principally in European, Asian, and North African lands. It includes, inter alia, works of legal argumentation and responsa, commentaries on Scripture and Talmud, philosophical treatises, and even works on scientific matters.

The language employed in this corpus contains many unique linguistic characteristics which differentiate it from other layers of Hebrew, and thus pose a great challenge for computational analysis. Heretofore no syntax-annotated corpus has existed for this layer of Hebrew, and existing parsers for modern Hebrew fall flat when applied to this corpus.

In order to fill this lacuna and pave the way for computational analysis of this sizeable corpus which comprises the foundation of Jewish culture and law, we have embarked upon a new project – the first of its kind – to annotate a representative set of post-Rabbinic historical Hebrew sentences as per the Universal Dependencies standard. We are pleased to announce the completion of the inaugural batch of this corpus, and to release the corpus to the community.

2 Existing Hebrew Corpora

The Hebrew language is currently represented in four UD corpora. The first of these corpora, UD HTB (Sade et al., 2018), building on the work of the original HTB (Winter et al., 2001), provides 6143 sentences (114K tokens) of modern Hebrew, taken from the newspaper *Ha'aretz*. This corpus laid the groundwork for application of the UD guidelines to Hebrew, with regard to dependency relations and segmentation of space-delimited tokens into syntactic words. Zeldes et al. (2022) recognized the need for a more diverse corpus and created a new corpus ("IAHLTwiki") of 5K sentences (140K tokens) from 39 Hebrew Wikipedia articles spanning 7 domains. They suggest adjustments to the conventions used in UD HTB's conventions for both segmentation and dependency relations.[2] This was followed by the IAHLTKnesset treebank (2800 sentences, 67K tokens), drawn from protocols held in the Israeli parliament between 1998 and 2022, further improving the diversity of the available corpora with the addition of spoken language (Goldin et al., 2024). All three of these treebanks cover only the most recent half-century of Hebrew. Swanson and Tyers (2022) began to rectify this with a corpus of Biblical Hebrew, consisting of 2.5K sentences from the books of Genesis, Exodus, Leviticus, and Ruth, totaling 62K tokens. The intervening two millennia of Hebrew, then, remained unspoken for.

[1] To be sure, there is a wide range of linguistic styles within this corpus, and this corpus can certainly be subdivided into multiple subdivisions (Goshen-Gottstein, 1985; Tènè, 1985). Nevertheless, on the whole, scholars have differentiated between four primary layers of Hebrew: Biblical Hebrew, Rabbinic Hebrew, post-Rabbinic historical Hebrew, and modern Hebrew, and it is this division which motivates our paper (Ben-Hayyim, 1985; Rabin, 1985).

[2] For maximum compatibility, we release our new corpus both in HTB format as well as IAHLT format.

Proceedings of the 23rd International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2025), pages 91–96
August 28-29, 2025 ©2025 Association for Computational Linguistics

3 The Present Corpus

The present corpus comprises over 11,000 words of text from a variety of post-Rabbinic Hebrew sources. Full details of the sources and sentence/word counts are provided in Table 1. Each sentence was initially annotated in terms of its syntactic functions and dependencies by one of our four linguistic experts (the first four authors of the present paper). Afterward, the four linguists convened together and critically reviewed each annotation, adjusting and honing the annotations to ensure both accuracy and consistency.[3]

4 Unique Syntactic Characteristics of Post-Rabbinic Historical Hebrew

As noted, from a linguistic perspective, the norms of post-Rabbinic historical Hebrew are quite different from that of modern Hebrew; hence the need for a new annotated corpus. In this section we survey three such differences.

4.1 Dislocated elements

The dislocated elements in our corpus are primarily cases of topicalization. In post-Rabbinic texts, dislocated topicalization is very common, much more so than in modern Hebrew. Thus, in IAHLTwiki there are 24 dislocated elements in the entire corpus, and only four of them are cases of topicalization (0.04% of the sentences), and in HTB there are 13 cases of dislocated topicalization (0.21%). In IAHLTknesset, which contains many transcriptions of spoken Hebrew, the frequency of dislocated topicalizations is a bit higher, yet still limited to one percent (29 sentences out of 2883). By contrast, in our corpus, there are 37 dislocated elements (11.5%) almost all of which are cases of topicalization.

To the limited extent that dislocated elements do occur in modern Hebrew in written form, they generally appear with a comma (or other punctuation mark) which indicates the boundary of the extraposed part. In post-Rabbinic historical Hebrew, there are generally no punctuation marks. Therefore, a syntax parser trained on the existing modern Hebrew corpora is likely to fail to locate the extraposed part and identify it as a dislocated element. Consider the following sentence:[4]

(1) Ḳaṭan ha-yodeaʿ lehitʿaṭef
small DET/SCONJ-knows to.wrap
aviṿ tsarikh liḳaḥ lo tsitsit
his.father must to.take to.him tsitsit
leḥankho
to.educate.him
‘A small boy who knows how to wrap himself [in a tallit] his father must buy him [a tallit with] tsitsit to educate him.’
[Sentence ID: 241]

The dislocated element, "a small boy who knows how to wrap himself," appears, due to the lack of punctuation, as though it were the subject of the sentence. In fact, however, the subject is "his father", meaning the father of that boy. This structure appears here, as in many other instances in our corpus, in place of a conditional clause ("If a boy knows how to wrap himself, his father must take. . . ").

Indeed, when running this sentence through the DictaBERT syntactic parser (Shmidman et al., 2024) – a syntax parser trained on modern Hebrew – it fell into this very trap. It labeled both *ḳatan* (a small boy) and *av* (father) as `nsubj`, each dependent on words within the main component of the sentence (*tsarikh* and *liḳaḥ*, respectively). This results in an illogical dependency parsing.

4.2 Causal clauses introduced by "*she-*" alone

Causal clauses in Hebrew begin with various connective words, including *ki*, *mi-pene she-*, *mishum she-*, and others. In rare cases in biblical language and more commonly in post-Rabbinic Hebrew, we find the prefix *she-* and its (originally Aramaic) equivalent *de-* used as a subordinating conjunction for causal clauses without a preceding connective word. The appearance of *she-/de-* without a connective word creates a structure which normally indicates a relative clause, and thus, in most such instances, the correct analysis of the sentence can only be determined based on semantics.

Such causal clauses appear in our corpus more than 15 times (more than 5% of sentences). By contrast, causal clauses with this syntax are highly irregular in modern Hebrew, and do not exist in the available annotated corpora of modern Hebrew.[5]

[3] Because the corpus comprises less than 20K words, we have not performed a train-dev-test split, as per `https://universaldependencies.org/release_checklist.html`; rather, we recommend testing via 10-fold cross validation.

[4] Example sentences in the linguistic discussions in this paper are presented in a format that maximizes readability, but which sometimes diverges from the UD segmentation (unless otherwise indicated). For the proper UD tokenizations and tagging of the example sentences, please reference the corpus using the provided sentence IDs.

[5] To be sure, other (non-causal) types of adverbial clauses

Work	Sentences	Words	Author	Author d.	Location	Description
Igeret orhot olam	13	1084	Abraham Farissol	1525	Italy	Geographic/cosmographic studies.
Sefer ha-hinukh	14	675	Unknown	~13th C.	Spain	Treatise on Biblical Commandments.
Rashi la-Torah	100	2161	Shlomo Yitzchaki	1105	France	Rashi’s Pentateuchal Commentary.
Shulhan arukh	80	4015	Joseph Karo	1575	Israel	15th century code of Jewish Law.
Sefer Maharil	18	569	Yaakov HaLevi Moelin	1427	Germany	Record of Ashkenazic Customs
Miscellaneous	92	3288	—	—	—	Eclectic collection of sentences from throughout post-rabbinic literature.
TOTAL	318	11802	—	—	—	—

Table 1: A summary of the texts included within our annotated post-Rabbinic historical Hebrew UD corpus.

Therefore, a syntactic analyzer of modern Hebrew has difficulty dealing with it. For example[6] (2):

(2) ṿe-en omrim kol yeme nisan tsidḳatkha
and-no saying all days.of Nisan Tsidkatkha
de-dino kemo teḥinah
SCONJ-its.status like supplication
‘And "Tsidkatkha" ("Your righteousness") is not recited [during] all the days of the month of Nisan, for its status is that of a supplication.’
[Sentence ID: 131]

We demonstrate the difficulty by running the sentence through the DictaBERT parser (Shmidman et al., 2024) (see Figure 1). In Figure 2 we present the correct analysis, as we have analyzed it in our corpus.

4.3 Conjunctions

In our corpus, we often find verbal elements within a single clause which are not of the same tense, yet are joined by coordinating conjunctions, as in (3), (4). Such a conjunction is expected between different sentences, but not within the same sentence. In modern Hebrew, such a conjunction is very rare, if not non-existent. We conducted extensive searches in the existing Hebrew treebanks and found no such conjunctions. For example:

(3) ha-holekh ba-derekh ṿe-higiaʿ
DET/SCONJ-walking in.the-way and-arrived
la-ʿir ṿe-rotseh lalun bah
to.the-city and-wishes to.lodge in.it
‘[One] who travels and arrived at a city and wishes to lodge therein.’
[Sentence ID: 295]

The sentence begins with a present participle, "ha-holekh" (Adler et al. 2008), and continues with a past tense verb "ve-higiaʿ".[7] These two words together, in coordination, comprise the root of the syntactic subject of the clause, and we would expect them to be of the same tense. Alternatively, the conjunction could have been replaced by a relative pronoun. The use of coordination here diverges from normative syntax of modern Hebrew.

(4) tsarikh leha’arikh be-ḥet shel eḥad ...
must to.lengthen in-Ḥ of EḤAD ...
ṿe-ya’arikh be-dalet shel eḥad
and-he.will.lengthen in-D of EḤAD
‘One must hold (i.e. tenuto) the h of "ehad" (=one) . . . and will hold the d of "ehad".’
[Sentence ID: 288]

The legal imperative in Hebrew can be expressed in several ways, e.g. by impersonal verb (see: Mor and Pat-El 2016) or future tense. We would expect to find a single mode in a given citation, but here we have a mixture of the two – impersonal verb ("tsarikh le-ha’arikh") and future tense ("ṿe-ya’arikh").

5 Annotation Decisions

5.1 UD tags specific to this corpus

In order to capture the linguistic complexities of this corpus, we have added a number of new features to the UD annotation. All the new features are materialized as subtypes of existing UD tags; thus the corpus remains valid for crosslingual comparison, as illustrated in Swanson et al. (2024).

5.1.1 part

The use of the Hebrew participle effects unique syntactic constructions. This is because, on the

are occasionally subordinated by a *she-* alone, and these do appear in those corpora, albeit very rarely. Regarding *she-*clauses in general, see Kogut (1937). *De-* does not appear in modern Hebrew.

[6] We have brought here a sentence that actually uses the Aramaic *de-* in the original text; however, because this sentence will be used to show the inability of the modern Hebrew parser to analyze such sentences, we adjusted to the Hebrew *she-* in order to give the parser a fighting chance.

[7] Regarding the double gloss of the Hebrew "ha" clitic at the start of the sentence: this clitic straddles the boundary between a definite article and a relativizer. In practice, in the corpus, any given instance of the clitic is specified as either SCONJ or DET, because only a single value can be selected.

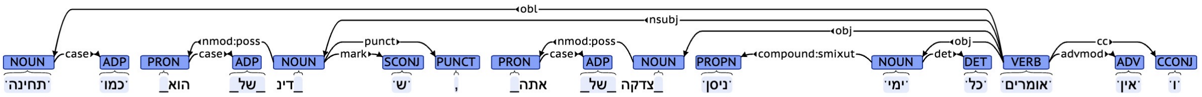

Figure 1: The sentence as analyzed by a syntactic parser trained on modern Hebrew (in right-to-left reading order). The causal clause is analyzed by the parser as two separate parts: its subject is parsed as the subject of the main sentence; its predicate is parsed as an oblique argument of the main verb; the subordinator '*she*-' is illogically tagged as a mark to the subject of the subordinate clause.

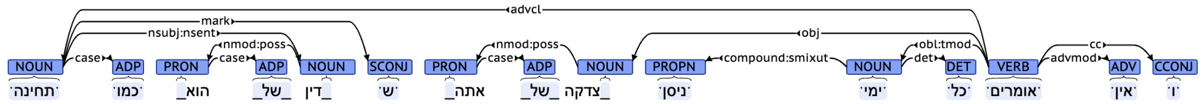

Figure 2: The sentence as annotated in our corpus by our linguistic experts.

one hand, it is a verbal form, but on the other hand, it can serve as a nominal or adjectival component. (Rosén 1956; Zewi and Reshef 2009; Sharvit 1980; Adler et al. 2008).

Consider the participle at the beginning of (3). The part-of-speech of "holekh" ("is walking"), as a participle, is VERB. In a sense, it is the root of a clause that complements an assumed nominal "one": "[one] who is walking in the way". Yet, the word in question is the subject of the sentence, and therefore tagged as nsubj.[8] Furthermore, the "ha" clitic attached to it normally serves as a definite article attaching to nominals.[9] That is, the obl "way" is, remarkably, a complement of a word with notable nominal morphosyntactic characteristics. Moreover, participles appear in our corpus in nominal positions other than nsubj; for instance, as the second member of a genitive construct (*smikhut* – a Hebrew construct that connects two nominals).[10]

In all of the aforementioned cases, the participle straddles the border between verb and noun, and an annotation that ignores this would be misleading and inconsistent. In order to bridge this gap and to properly represent the complexity of these participle forms, we have labeled their own dependency relations according to their nominal syntactic function, while adding the part subtype to their dependents that function as verbal complements and thus reflect the verbal character of the participle.

5.1.2 conj:push

Many times in post-Rabbinic Hebrew we find a series of coordinating conjunctions which are not equal in their syntactic value. For instance, coordinating conjunctions often appear where we would expect a subordinating conjunction, such that we end up with a case of nested coordination (Universal Dependencies). In order to capture this nuance, we have added a push subtype; specifically, in a structure of type (A, B), C, the push subtype is specified for B.

5.2 Tokenization of negative particles

In negative particles of type "*en* ('no') + personal pronoun", the pronominal component sometimes serves as the subject of the sentence (see e.g. (5)). In other cases, however, it simply negates a sentence that has an explicit subject (with which the pronominal component agrees – see e.g. (6)). In the former case, the particle contains two syntactic words; in the latter, it contains one. In existing modern Hebrew treebanks, (see Section 2), such particles are always tagged as a single word. In the biblical Hebrew treebank (ibid.), they are always split.[11] We have chosen to differentiate between the cases: when the sentence contains no other explicit subject, we split the token into the negating particle *en*, which receives an advmod dependency relation, and the corresponding pronoun, which receives an nsubj dependency relation. When the

[8] Here we follow the UD guidelines for Hebrew (Universal Dependencies Contributors, 2024), under which such participles are tagged nsubj and not csubj.

[9] See footnote 7.

[10] See sentence 173 in the tagged corpus for such an example. See also sentence 127, in which a participle has both obl and case dependents.

[11] This seems to follow the approach that pronouns in biblical Hebrew do not normally serve as copulae (for discussion see Joüon and Muraoka 2011, §102*k* and §154*i*). In this approach, all sentences with an explicit subject and a negative particle with a pronominal suffix are naturally analyzed as *casus pendens* – that is, what appears to be the explicit subject is in fact a dislocated element, and the negative particle contains the subject.

subject appears, we segment as in the modern Hebrew treebanks. An example of our segmentation (with selected morphological features added for clarity) for each kind of sentence appears below.

(5) en-i rotseh lehinaḳem mi-Ḳayin ʿakhshaṿ
no-I want to.be.avenged from-Cain now
‘I do not want to take revenge on Cain now.’

[Sentence ID: 153]

(6) ḳṭanah **enah** yekholah laʿaśot
small.FSG **no.3MSG** able to.make
shaliaḥ
messenger
‘A minor girl cannot appoint a proxy.’

[Sentence ID: 313]

6 Conclusion

We have prepared the first UD-tagged corpus of historical post-Rabbinic Hebrew, containing over 11,000 words across multiple genres and time periods. We are pleased to release this corpus to the public. The corpus is released under a Creative Commons Attribution 4.0 International License (CC BY 4.0). The corpus is available now on github,[12] and we hereby submit the corpus to the upcoming UD release as a new treebank within the *heb* language.

Acknowledgements

This work has been funded by the European Union (ERC, MiDRASH, Project No. 101071829; Principal investigators: Nachum Dershowitz, Tel-Aviv University; Judith Olszowy-Schlanger, EPHE-PSL; Avi Shmidman, Bar-Ilan University, and Daniel Stoekl Ben Ezra, EPHE-PSL), for which we are grateful. Views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

Meni Adler, Yael Dahan Netzer, Yoav Goldberg, David Gabay, and Michael Elhadad. 2008. Tagging a Hebrew corpus: the case of participles. In *LREC*.

Zeev Ben-Hayyim. 1985. The problems of the unity of the Hebrew language throughout its history and its periodization. In Moshe Bar-Asher, editor, *Language Studies*, volume 1. Magnes.

Gili Goldin, Nick Howell, Noam Ordan, Ella Rabinovich, and Shuly Wintner. 2024. The Knesset corpus: An annotated corpus of Hebrew parliamentary proceedings.

Moshe Goshen-Gottstein. 1985. Corpus, genre and the unity of Hebrew – aspects of conceptualization and methodology. In Moshe Bar-Asher, editor, *Language Studies*, volume 1. Magnes.

Paul Joüon and Takamitsu Muraoka. 2011. *A Grammar of Biblical Hebrew*. Gregorian and Biblical Press, Roma.

Simcha Kogut. 1937. *The Complex Sentence in Sefer Hasidim*. Ph.d. dissertation, Hebrew University of Jerusalem. See especially pp. 204–207.

Uri Mor and Na’ama Pat-El. 2016. The development of predicates with prepositional subjects in Hebrew. *Journal of Semitic Studies*, 61(2).

Chaim Rabin. 1985. Periods of the Hebrew language. In Moshe Bar-Asher, editor, *Language Studies*, volume 1. Magnes.

Haiim B. Rosén. 1956. *Our Hebrew: Its Character in Light of Linguistic Methods*. Am Oved, Tel Aviv.

Shoval Sade, Amit Seker, and Reut Tsarfaty. 2018. The Hebrew Universal Dependency treebank: Past present and future. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 133–143, Brussels, Belgium. Association for Computational Linguistics.

Shimon Sharvit. 1980. The tense system in Mishnaic Hebrew. In G. Sarfatti, P. Artzi, H. Y. Greenfield, and M. Z. Kaddari, editors, *Studies in Hebrew and Semitic Languages Dedicated to the Memory of Prof. Yechezkel Kutscher*, pages 110–125. Bar-Ilan University, Ramat Gan.

Shaltiel Shmidman, Avi Shmidman, Moshe Koppel, and Reut Tsarfaty. 2024. MRL parsing without tears: The case of Hebrew. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4537–4550, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

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.

Daniel G Swanson, Bryce D Bussert, and Francis Tyers. 2024. Producing a parallel universal dependencies treebank of Ancient Hebrew and Ancient Greek via cross-lingual projection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13074–13078.

[12] https://github.com/ERC-Midrash/UD_Hebrew-PostRab

David Tènè. 1985. Historical identity and unity of Hebrew and the division of its history into periods. In Moshe Bar-Asher, editor, *Language Studies*, volume 1. Magnes.

Universal Dependencies. Universal dependencies - nested coordination. Accessed: 2025-04-07.

Universal Dependencies Contributors. 2024. UD for Hebrew. Accessed: 2025-06-12.

Yoad Winter, Alon Altman, K. Sima'an, Alon Itai, and Noam Nativ. 2001. Building a tree-bank of modern hebrew text.

Amir Zeldes, Nick Howell, Noam Ordan, and Yifat Ben Moshe. 2022. A second wave of UD Hebrew treebanking and cross-domain parsing.

Tamar Zewi and Yael Reshef. 2009. The active participle and temporal expression in Hebrew. *Leshonenu*, 71(3–4):315–344.