

A Second Wave of UD Hebrew Treebanking and Cross-Domain Parsing

Amir Zeldes

Georgetown University
amir.zeldes@georgetown.edu

Nick Howell

IAHLT
nlhowell@gmail.com

Noam Ordan

IAHLT
noam.ordan@gmail.com

Yifat Ben Moshe

IAHLT
yifat1811@gmail.com

Abstract

Foundational Hebrew NLP tasks such as segmentation, tagging and parsing, have relied to date on various versions of the Hebrew Treebank (HTB, [Sima'an et al. 2001](#)). However, the data in HTB, a single-source newswire corpus, is now over 30 years old, and does not cover many aspects of contemporary Hebrew on the web. This paper presents a new, freely available UD treebank of Hebrew stratified from a range of topics selected from Hebrew Wikipedia. In addition to introducing the corpus and evaluating the quality of its annotations, we deploy automatic validation tools based on [grew](#) ([Guillaume, 2021](#)), and conduct the first cross-domain parsing experiments in Hebrew. We obtain new state-of-the-art (SOTA) results on UD NLP tasks, using a combination of the latest language modelling and some incremental improvements to existing transformer based approaches. We also release a new version of the UD HTB matching annotation scheme updates from our new corpus.

1 Introduction

Treebanks (TBs) form a fundamental resource for NLP and computational linguistics research: they provide high quality annotated data for tokenization, sentence splitting, POS tagging and syntactic/semantic relation extraction. For Morphologically Rich Languages (MRLs, [Seddah et al., 2014](#)), high quality morphosyntactically annotated data is particularly crucial, since basic search on a word-matching level cannot work before morphological segmentation has occurred. In languages such as Hebrew, where vowels are not well represented in the script, the need for reliable morphosyntactic disambiguation is particularly strong, since string-level ambiguity is very frequent and high. This is demonstrated by the following frequently cited example from [Adler and Elhadad \(2006\)](#), which has a large number of possible analyses for just a four character sequence (note Hebrew is right-to-left).

- (1) **בצלם**
<b.cl.m> be.cil.am - in.shadow.their
<b.clm> (be./b.a.)celem - in.(a/the).image
<b.clm> (be./b.a.)calam - in.(a/the).photographer
<bcl.m> bcal.am - onion.their
<bclm> becelem - Betzelem (organization)

When such sequences contain multiple sub-tokens (e.g. ‘in’ and ‘the’ within ‘in.the.image’), we follow Universal Dependencies¹ terminology in referring to the larger unit as a Multi-Word Token (MWT), and the sub-parts, each of which carries a separate part-of-speech, as tokens.

Although a treebank for Hebrew was already created by the MILA Institute ([Sima'an et al. 2001](#), hence HTB), subsequently converted to dependencies ([Goldberg and Elhadad, 2009](#)) and finally to the more recent Universal Dependencies framework (see [Sade et al. 2018](#), hence UD-HTB), two major concerns motivate the current work in creating a new UD treebank for Hebrew. The first is the age of the data: HTB texts were taken from 1990–91 issues of a single newspaper, Ha’aretz, as illustrated in (2), describing the introduction of computers to an office.

- (2) *ba-axrona huxnesu ma'arexet maxshev vetoxnot le-kol ha'anafim*
‘Recently a computer system and software were introduced into all branches’

This is one of only a dozen examples of the word ‘computer’ in the data, which predates mainstream Web 1.0 times, mentions no cellular phones, the Internet, or a variety of countries established after 1990, including the EU. Previous work has shown that NLP systems retain strong lexical biases mirroring both period and author demographics ([Shah et al., 2020](#)), which in the case of HTB reflect primarily Israeli Ha’aretz journalists from 1990.

A second concern beyond the period and author demographics is genre. HTB is a news corpus, and

¹<https://universaldependencies.org/>

as such reflects formal journalistic writing, which focuses on describing prominent political events, sports news, and reported speech, usually in the past tense, but under-represents expository text, academic language, and colloquial spellings (which are much more variable in Hebrew than in English). The importance of genre diversity in training data has often been noted (Zeldes and Simonson, 2016; Müller-Eberstein et al., 2021), but without a second genre to test on, we simply do not know the extent to which fundamental Hebrew NLP performance degrades outside of HTB’s language.

In this paper, we attempt to broaden the range of data available for Hebrew by creating and evaluating a new, freely available (CC-BY-SA license), gold standard corpus using contemporary Wikipedia data. Although Wikipedia’s language is also relatively formal, it differs substantially from newspaper reporting, and is more contemporary than HTB, covering a broad range of topics, while being available under an open license.² Our main contributions are:

1. A new TB of Hebrew Wikipedia data from several domains annotated for all UD layers, including morphological segmentation and features, POS tags and dependency trees, and report first UD Hebrew agreement scores
2. New SOTA results on the standard benchmark for Hebrew segmentation, tagging and parsing
3. The first cross-corpus evaluation of out-of-domain (OOD) Hebrew NLP across all UD tasks; we also perform error analysis indicating some issues with previous benchmark data
4. We release all code and trained models for the tools evaluated in Section 4, including new models for all tasks using the popular Stanza and Trankit libraries, as well as a new SOTA library tailored for Hebrew NLP

2 Previous work

In terms of material, there is only one existing TB of modern Hebrew prior to our work (Sima’an et al., 2001) based on 1990–91 issues of the newspaper Ha’aretz. However, there are multiple versions of

²We also note the current development of UD resources for Biblical Hebrew (https://github.com/UniversalDependencies/UD_Ancient_Hebrew-PTNK), however we concur with UD in classifying ancient varieties as separate languages.

this dataset, leading to some confusion. The original TB contained constituent trees, as well as word-internal segmentation and treebank-specific POS tags. This dataset was converted into dependencies first by Goldberg and Elhadad (2009) using a custom scheme, and later to an early version of UD by Tsarfaty (2013). The current UD HTB, using an older version of UD V2 guidelines which became invalid in 2018, is described in Sade et al. (2018), and is used below for evaluation using legacy tokenization. An updated version of this dataset released in this paper is based on the 2018 version.

This work represents the second treebank ever produced for modern Hebrew, and the first cross-domain NLP evaluation of dependency parsing for the language. We briefly survey the state-of-the-art in UD Hebrew NLP in Section 4.

3 The new corpus

3.1 Contents

Our new corpus, IAHLTWIKI,³ created by the non-profit Israeli Association of Human Language Technology (IAHLT), contains 5K sentences taken from Wikipedia, which were annotated by a team of 6 annotators, who all have either an undergraduate or graduate degree in Linguistics and a robust knowledge of Hebrew syntax, over 6 months using the Grew-Arborator tool (Guibon et al., 2020). During annotator training, a member of the UD core group was consulted whenever questions arose in a series of meetings. Data was sampled from 7 topic categories in order to increase lexical and structural diversity: biographies, events, places, health, legal, finance and miscellaneous (see Table 1). Inclusion of the last category is intended to introduce some random topic variation into the data, while the former categories were selected in consultation with a consortium of Israeli industry partners based on high interest applications for information extraction regarding people, places, healthcare, etc.

Rather than sampling random sentences, full Wikipedia entries were selected in order to allow for future annotation projects at the document level, such as salient entities, coreference resolution, or other types discourse annotation. Domain data varies somewhat in size, though most domains cover roughly 20-30K tokens, with the exception of *finance* and *event*, to which we devoted less space

³https://github.com/UniversalDependencies/UD_Hebrew-IAHLTWiki

domain	documents	tokens	sentences
<i>bio</i>	5	21,963	754
<i>event</i>	4	16,202	580
<i>finance</i>	5	8,723	299
<i>health</i>	8	20,927	824
<i>law</i>	6	22,916	788
<i>place</i>	5	22,323	829
<i>misc</i>	6	27,895	965
total	39	140,949	5,039

Table 1: Domains in the IAHLTWIKI corpus.

based on our priorities. Together they were given the same total amount of data as a single domain.

To illustrate some of the lexical differences between the older HTB news data and IAHLTWIKI, Table 2 shows words which are over/under represented in each dataset compared to the other (ratio skewed towards wiki in shades of blue, or HTB in red), taken from the top 50 words sorted by the frequency ratio, and constrained to include a maximum of 3 words not appearing in the other dataset. Some of the items on the left relate to period: ‘2018’ is obviously not discussed in HTB, but neither is ‘LGBT’, which was a very new term in 1990 and not yet translated to Hebrew. Other items relate to domain: inclusion of medical articles leads to 33 occurrences of ‘Penicillin’ (0 in HTB) and 10 times more mentions of ‘disease’. Other differences are stylistic: the spelling *הייתה* for the 3rd person singular feminine *hayta* ‘was’ is very common colloquially, but ‘sub-standard’ for Ha’aretz, which always spells it with one ‘y’: *היתה*.

On the opposite side of the table, we also see period effects: ‘Soviet’ or ‘Kuwait’, which played important roles in world news in 1990; but also common items missing from Wikipedia, such as ‘day before yesterday’, a relative time term unlikely to appear in encyclopedic text, except perhaps in a quotation. Other underrepresented items interact with genre, such as ‘week’ (very common in narratives, less in exposition) or ‘but’ (more typical of argumentative writing).

Overall our corpus contains around 14K lexical items, nearly 7K of which are missing from HTB, not only due to its topics, but also possibly due to the fact that Wikipedia is authored by diverse volunteers, and is not guided by a newspaper’s stringent style policy and proofreaders. Sentences in the corpus also trend longer, which may be of value for parser performance on longer sentences: $M = 25.02$ and $SD = 14.27$ tokens in HTB, compared to

$M = 27.97$ and $SD = 15.79$ in IAHLTWIKI.

3.2 Changes from HTB

The previous version of UD HTB has been invalid based on the official UD validation page since 2018, meaning several schema changes were needed to conform to the latest UD standards. To make comparison of the datasets and joint training possible, we release and evaluate a newly revised version of UD HTB which is valid by UD 2.10 release standards. This new version was initially created via automatic scripts using the DepEdit Python library (Peng and Zeldes, 2018) followed by manual post editing and validation, and nearly doubles the total amount of UD data for Hebrew, (over 11K sentences), allowing for cross-domain experimentation.

Tokenization Besides correcting thousands of HTB errors and improving consistency, we introduce a major change to tokenization, which ensures that multiword tokens always correspond to the concatenation of their sub-tokens. In particular, we remove inserted possessive של [šel] ‘of’ between nouns and their clitic possessives (as in (3) for the MWT *ביתו* [beit-o] ‘his house’), object marking *את* [et] (accusative marker, as in (4) for *ראייתה* [re’iti-ha] ‘(I) saw her’) and orthographically unexpressed articles (as in (5) for *בבית* [ba-bait] ‘in the house’).

- (3) a. Old: *בית-של-הוא* [bait _shel _hu]
b. New: *בית-ו* [beit o]
- (4) a. Old: *ראיתי-את-היא* [ra’iti _et _hi]
b. New: *ראיתי-ה* [re’iti ha]
- (5) a. Old: *ב-ה-בית* [be _ha_ bait]
b. New: *ב-בית* [ba bait]

In (3), the former HTB tokenization inserted the word *shel* ‘of’ surrounded by underscores before clitic possessors, as though the example reads ‘house_of_him’, rather than ‘his house’. This complicates tokenization and introduces an unnecessary inconsistency with related languages in UD, such as UD Arabic, which does not insert an unexpressed preposition in the same construction, nor object markers in cases like (4).

The last case in (5), i.e. removal of zero articles (which speakers can reconstruct from context while reading Hebrew but are not trivial to predict) is the only change resulting in a loss of information, and we therefore replace these by a standard UD morphological feature `Definite=Def`. As a result,

Under-represented in HTB					Under-represented in Wikipedia data				
word	translation	Wiki	HTB	ratio	word	translation	Wiki	HTB	ratio
פניצילינים	<i>Penicillins</i>	33	0	0	תיירות	<i>tourism</i>	0	31	inf
2018	<i>2018</i>	22	0	0	שלשום	<i>day before yesterday</i>	0	29	inf
להט"ב	<i>LGBT</i>	3	0	0	כוויית	<i>Kuwait</i>	0	26	inf
הייתה	<i>was (3.sg.F)</i>	141	1	0.007	שבוע	<i>week</i>	8	110	13.75
אלבום	<i>album</i>	183	3	0.016	משחק	<i>game</i>	9	115	12.777
שירים	<i>songs</i>	127	6	0.047	אבל	<i>but</i>	24	207	8.625
מחלה	<i>disease</i>	35	3	0.085	משטרה	<i>police</i>	20	149	7.45
ערבית	<i>Arabic</i>	30	4	0.133	ח"כ	<i>member of parliament</i>	7	49	7
יוונית	<i>Greek</i>	22	3	0.136	דולר	<i>dollar</i>	13	89	6.846
אשראי	<i>credit</i>	11	2	0.181	מועצות	<i>Soviet</i>	10	37	3.7

Table 2: Words with striking frequency differences sorted by ratio of frequency in HTB vs. the Wiki data. Ratio for words overrepresented in Wikipedia is shaded blue, and for HTB in red.

conversion between the new and old TB tokenization is deterministic in both directions, but the new tokenization style is both simpler and matches practices in related languages, most notably Arabic.

Dependencies Some custom relation sub-types, which were predictable from the words they connect, have been removed in the new scheme, such as `case:acc` and `case:gen` for object and possessive markers, or `mark:q` for the question markers.

3.3 Validation

The standard UD toolkit⁴ contains a validation framework, but at the language-specific level it is quite limited, largely checking feature-POS combination sanity and permitted relations or auxiliaries. Ideally, each enforceable provision in language specific guidelines should be captured in machine-readable format, a vision we attempt to implement using *grewv*, an extension of the Grew “graph rewriting for NLP”, a search and transformation engine for decorated graphs, originally designed for corpus exploration (Guillaume, 2021). Grew allows rigorous definition of graph patterns, incl. quantification and negation, which enables e.g. searching for verbs with no subject (Figure 1). We created a tool around *grew*, dubbed *grewv* (for “grew validator”), which uses *grew* expressions to describe *non-conformant* trees and generate an error report in human- and machine-readable formats.

As we encountered new phenomena in the data, annotators worked together to define guidelines, but also rules, consisting of a graph matching definition, a short name, a long-form message template, and an error level: ‘error’ or ‘warning’. Rule templates support placeholders for tree-specific infor-

mation, such as token numbers. While errors result in validation failure, warnings can be dismissed, indicating that an annotator has reviewed and approved the data. An error-level message and pattern are shown in (6), which prohibits tokens from bearing the `cc` label for coordinating conjunctions if they are not the child of a coordinated token, with one Hebrew lexical exception for the expression `בין ובין` ‘both ... and’.

- (6) **error:** *Token* {matching[nodes][X]} has a *cc* child (token {matching[nodes][Y]}), but is neither *conj*, *parataxis* nor *root*.
pattern: { Y[lemma<>"בין"]; X-[cc]->Y; }
without { *-[conj|root|parataxis]-> X; }

We integrated *grewv* into all stages of the project: continuous validation is used on the ruleset and at annotation time, and the treebank is passively scanned for trees requiring correction on updates. Changes to the ruleset are submitted, discussed, and reviewed over a mailing list, and proposed changes trigger an analysis job on our continuous integration system. Each change is checked for syntax errors and illustrative examples of trees which are supposed to pass and fail. Finally, a random subsample of outputs is inspected by an annotator to confirm that the rule works as expected.

Confirmed changes are automatically deployed to our Arborator-Grew annotation instances; we have made modifications to track and report validation errors and warnings, and annotations cannot be marked ‘final’ until they have passed validation. As rules evolved, the treebank is passively scanned for ‘stale’ trees: annotations which passed validation at the time they were made, but now require correction in order to pass the latest ruleset. Periodically batches of stale trees are prepared for freshening,

⁴<http://github.com/UniversalDependencies/tools>

```

1 % Search for verbs without subject
2
3 % basic pattern: a verb with some constraints on VerbForm and Mood
4 pattern { V [upos="VERB", VerbForm <= Ger|Inf|Part, Mood <= Imp] }
5
6 % first negative pattern: there is no "subject" node
7 without { V -[nsubj|csubj|nsubj:pass]-> S}
8
9 % second negative pattern: the verb if not a dependent of a relation "cop", "aux", ...
10 without { N -[cop|aux|aux:pass|conj]-> V}

```

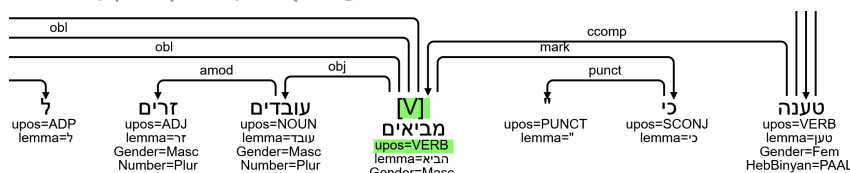


Figure 1: A Grew search in HTB for subjectless verbs and a result.

to ensure the treebank asymptotically approaches a consistent annotation level.

3.4 Inter-annotator agreement (IAA)

Table 3 gives aligned token accuracy using the official CoNLL scorer (Words) and Cohen’s κ for each annotation for ~ 440 double annotated sentences ($\sim 12.6K$ tokens) before adjudication and grew validation. As this is the first Hebrew corpus annotated natively in UD, these are, to the best of our knowledge, the first reported IAA scores on full UD annotation for the language. Due to the nature of κ , features beyond segmentation are computable only for identically tokenized words, however due to the high segmentation score this is negligible.

Words	Lemma	UPOS	FEATS	Head	Deprel	Misc
99.1%	98.5	94.8	90.9	95.8	95.7	95.6

Table 3: Word segmentation accuracy, and Cohen’s kappa for each annotation layer. Misc stands for miscellaneous optional UD annotations, including e.g. CorrectForm for typos.

The metrics indicate a very high level of agreement, with lowest scores on FEATS, where agreement holds only if all features agree (Gender, Number, Tense etc.); for κ -scores on individual morphological features, see the Appendix.

4 In-domain and cross-corpus NLP

In this section we present a series of experiments on cross-corpus segmentation, tagging, lemmatization and parsing, which is made possible for the first time by the release of non-newswire treebank data. Several systems exist for UD parsing of Hebrew: in this section we compare two popular neural off the shelf, end-to-end UD systems: Stanza (Qi et al., 2020) and Trankit (Nguyen et al., 2021), the latter of which is the SOTA for Hebrew parsing; and the

current SOTA system for non-concatenative Hebrew segmentation and tagging (Seker et al., 2022), which is the only one using a pre-trained Hebrew Transformer introduced in the same paper, AlephBERT, but which does not cover UD parsing based on the paper.⁵ We release and describe our own system for these tasks, called HebPipe,⁶ below, which is a pipeline system building on the best existing systems with some incremental improvements.

Segmentation All previous work on UD Hebrew parsing has been limited by the use of non-concatenative tokenization schemes. However with the conversion to the concatenative tokenization described in Section 3.2, we are able to take advantage of high accuracy characterwise segmentation approaches using binary classification (each character either begins a new segment or not). The previous characterwise SOTA system, RFTokenizer (Zeldes, 2018) is based on lexical features, such as looking up possible POS tags of MWT substrings around a segmentation point, and an XGBoost classifier, and has been applied to other morphologically rich languages, such as Arabic and Coptic (Zeldes and Schroeder, 2016); however it does not utilize pretrained transformers, which do not generally encode character level information (AlephBERT is word-piece based). Seker et al. (2022, 51) also mention that it is counterintuitive to use a word level model as input to character level tasks, but note that despite this their approach still achieves a high segmentation score.

To improve the SOTA in segmentation, we combine the characterwise RFTokenizer classifier with

⁵ Although the AlephBERT model has been released, code from the paper does not enable re-running training of the NLP components; we therefore report the scores from the paper below.

⁶ <https://github.com/amir-zeldes/HebPipe/>

the transformer from [Seker et al. \(2022\)](#) operating at the MWT word level. We train an LSTM using AlephBERT to predict *whether* each word contains prefixes requiring segmentation, suffixes, or both, thereby targeting an MWT-level property. We then feed these predictions to RFTokenizer as a feature, an approach not previously applied to the problem to the best of our knowledge. To avoid overfitting, we reserve half of the training data and the development set for the AlephBERT LSTM training, and the remaining half for the XGBoost classifier, which is fed the features described in [Zeldes \(2018\)](#) as well as the AlephBERT predictions (code attached to this paper).

Tagging For POS tagging and features we employ a simple LSTM sequence tagger using the same pre-trained AlephBERT embeddings. We observe in development data that some tokens are ambiguous regarding part of speech depending on whether they are prefixes or suffixes: for example, the letter vav ⟨w⟩ can stand for the word ‘and’, pronounced [ve] and tagged CCONJ, at the beginning of a MWT, or the word ‘him/his’, a pronoun pronounced [o] at the end of a MWT. If tokens are tagged as a naive sequence, these are occasionally mistagged in ambiguous contexts, for example:

(7) בית ו ארמון ⟨byt w `rmwn⟩

If the first two tokens are spelled together (ביתו), then the example is a predicative verbless sentence (‘his house is a palace’), and the letter ‘w’ is a pronoun (PRON). However if the last two tokens are spelled together (וארמון), then ‘w’ is a coordinating conjunction (CCONJ), meaning ‘and’, and the translation would be ‘a house and a palace’. Other errors can occur similarly, for example between an initial article *ha-* ‘the’ spelled the same as the suffix feminine possessive *-a* ‘her’.

To make the model aware of such cases, we concatenate a dense embedding using BIES notation to each token vector, indicating whether it begins a MWT (B), ends one (E), is inside one (I) or represents a single-token word (S). Using concatenative tokenization, we observe a degradation of -0.32% in end-to-end tagging accuracy on HTB by ablating this feature, which amounts to an error reduction of ~12%. Although this improvement is minor, it is not computationally complex, and we are hopeful that it could be a useful approach for other MRLs, such as Arabic. Tagger predictions are also fed into Stanza’s lemmatizer to produce our lemmatization.

Parsing For dependency parsing we rely on the prevalent biaffine approach proposed by ([Dozat and Manning, 2017](#)), using the implementation in DiaParser ([Attardi et al., 2021](#)), which adds intermediate transformer layer inputs to the transition classifier. We again rely on AlephBERT to represent input embeddings.

We use the default hyperparameters for all components described above and do not perform hyperparameter optimization (see appendix for details).

In-domain results Table 4 compares performance for the systems listed above. Results are computed using the official CoNLL scorer from the 2018 shared task on UD parsing, and with the exception of using gold sentence splits to match prior work on Hebrew, all numbers reflect realistic, end-to-end performance from plain text input, using the MWT F1 score for morphological segmentation, and AligndAcc for all other metrics. All numbers have been reproduced by retraining each system, except for [Seker et al.](#)’s system, for which trainable code is not provided. For HTB we report results using the concatenative tokenization, as well as with the older non-concatenative tokenization (in brackets) for comparability with previous work, with the official train-dev-test splits maintained. The splits for IAHLTWIKI were established by stratified sampling evenly across the domains described in Section 3.1.

As the table shows, our approach achieves new SOTA scores for segmentation (MWT F1), lemmatization, parsing (UAS and LAS) as well as morphologically and lexically informed combined UD scores (CLAS, MLAS, BLEX) on HTB.⁷ For tagging, the result is very close – given that both Trankit and our system use a torch-based, transformer-driven sequence tagger, it appears that either Trankit’s underlying XLM-RoBERTa model ([Conneau et al., 2020](#)) is superior to AlephBERT here, or we are looking at minor, random chance differences. Overall, the most notable advances in score are for segmentation – likely due to stacking RFTokenizer and AlephBERT, since the next best score is [Seker et al.](#)’s pure AlephBERT system – and for parsing, likely due to the use of a language specific BERT, which has not been reported on in previous papers for Hebrew. Here the difference to Trankit’s XLM-RoBERTa result is very noticeable (+6.6 points), however we note that this folds in

⁷See <https://universaldependencies.org/conll18/evaluation.html> for details on these metrics

	train on HTB → test on HTB				train on Wiki → test on Wiki		
	Stanza	Trankit	Seker et al.	This paper	Stanza	Trankit	This paper
MWT F1	93.97 (92.82)	97.27 (96.04)	– (98.20)*	98.81 (98.37)	92.87	94.64	98.78
POS	96.99 (97.12)	97.46 (97.63)	– (96.20)*	97.34 (97.40)	95.76	96.98	97.27
FEATS	95.45 (95.65)	85.95 (85.65)	– (93.05)*	91.68 (92.52)	89.23	90.69	91.06
AllTags	94.64 (94.85)	85.20 (84.90)	–	91.06 (91.92)	88.24	89.58	90.30
Lemma	96.63 (96.89)	96.69 (97.06)	–	97.52 (97.58)	97.51	94.70	97.49
UAS	85.62 (85.78)	85.46 (85.60)	–	91.90 (92.07)	82.96	89.81	92.19
LAS	82.67 (82.88)	82.82 (83.01)	–	89.42 (89.65)	80.22	87.65	90.01
CLAS	76.01 (75.93)	83.96 (83.88)	–	84.48 (84.48)	73.16	83.96	86.16
MLAS	70.63 (68.97)	63.00 (60.20)	–	72.24 (72.24)	59.64	69.35	72.37
BLEX	72.71 (72.65)	79.80 (79.87)	–	80.99 (80.99)	70.66	76.03	82.56

Table 4: In-domain UD NLP performance for both datasets. Figures in brackets are for the older, non-concatenative tokenization. All numbers are end-to-end from plain text. A “*” indicates numbers from the cited paper, other numbers are reproduced by the authors.

gains from the superior segmentation quality, since all scores are end-to-end from plain text sentences.

For IAHLTWIKI the results favor our approach even more, with the sole exception of lemmatization. However here our approach in fact uses Stanza itself, meaning the slightly lower score is probably due to random initialization differences. Although the absolute numbers for IAHLTWIKI are sometimes lower than for HTB’s scores, we note that the Wiki data may not only be more challenging, but the use of the simplified tokenization shifts task difficulties: based on HTB, segmentation accuracy is higher across the board when using the simpler tokenization; however due to the need to represent orthographically unexpressed articles using morphological features and the lack of easy-to-tag inserted pseudo-tokens, metrics involving morphology may become lower.

Cross-domain results In order to assess how reliably our results generalize to unseen data from a different source, we test the trained models from Table 4 on data from the ‘other’ corpus, using the same partitions (training still relies on each source corpus ‘train’ partition, and evaluation is on the same ‘test’ partition used in the previous table). For this cross-corpus evaluation, we report scores only on the new tokenization, since no previous scores are available for comparison for this setting, and systems are unable to learn non-concatenative MWT expansion from the new Wiki dataset. Table 5 shows the results for Stanza, Trankit and our system, HebPipe.

As the table shows, HebPipe achieves the best cross-corpus generalization, primarily due to large gains in MWT F1, which reduce cascading errors in downstream tasks. Segmentation degradation

is only around 0.25%, meaning POS tagging and lemmatization are mainly vulnerable to OOV items, which are mitigated by AlephBERT’s lexical coverage. Stanza and Trankit both suffer only a little over 1–2% tagging degradation, but for parsing metrics degradation is much more substantial (around 5–10% LAS on HTB).

Across the board, degradation is more substantial when predicting on HTB and especially on dependency metrics, which suggests that HTB contains more constructions not represented in the Wiki data. However, our error analysis suggests that a substantial portion of parsing degradation is owing to errors in the original HTB’s conversion into UD (see Section 5 below). Overall, these results indicate that OOD performance on a second genre of formal, written Hebrew is fairly reliable for segmentation and tagging, but less so for parsing, and that the segmentation approach taken in this paper is particularly robust, possibly due to XGBoost’s well known resistance to overfitting. We stress however that these results do not yet indicate performance quality on informal text types.

Joint training In response to reviewer feedback in the rebuttal period, we were able to train joint models for each component, using both the HTB and IAHLTWIKI train set for training and the joint dev sets for early stopping via simple concatenation. Table 6 gives the scores for the jointly trained model. While we do not necessarily recommend using a joint model due to possible inconsistencies between the datasets which have not yet been resolved (see Section 5 below), the numbers indicate that the model is able to robustly deal with both test sets, with neither substantial gains nor degradation over in-domain training numbers. Importantly, the

	train on Wiki → test on HTB				train on HTB → test on Wiki			
	Stanza	Trankit	This paper	Δ	Stanza	Trankit	This paper	Δ
MWT F1	91.79	92.24	98.59	-0.22	91.79	93.03	98.51	-0.27
POS	94.09	94.27	95.25	-2.09	94.09	95.02	96.41	-0.86
FEATS	83.35	81.19	88.02	-3.66	83.35	76.50	90.06	-1.00
AllTags	82.12	79.72	86.97	-4.09	82.12	75.10	88.94	-1.36
Lemma	95.05	90.96	97.31	-0.21	95.05	96.41	97.75	0.26
UAS	76.50	82.55	86.05	-5.85	76.50	82.32	90.37	-1.82
LAS	72.26	78.22	82.14	-7.28	72.26	78.34	87.35	-2.66
CLAS	63.65	72.62	76.29	-8.19	63.65	77.60	83.02	-3.14
MLAS	45.19	47.25	60.22	-12.02	45.19	45.10	67.90	-4.47
BLEX	58.63	60.88	72.86	-8.13	58.63	72.88	79.73	-2.83

Table 5: Out-of-domain UD NLP performance, training on one corpus and testing on the other, all numbers end-to-end from plain text. Δ gives the difference between HebPipe in-domain and out-of-domain performance.

model clearly outperforms cross-domain results, suggesting that it may indeed be overall the most robust choice for totally unseen data. We expect that further harmonization of the two datasets will increase the usefulness of joint training.

	HTB	Wiki		HTB	Wiki
MWT F1	98.78	98.65	UAS	91.11	92.65
POS	97.14	97.29	LAS	88.29	90.36
FEATS	91.02	90.93	CLAS	83.27	86.65
AllTags	90.39	90.11	MLAS	70.57	72.37
Lemmas	97.65	98.15	BLEX	80.13	84.03

Table 6: UD NLP performance for a joint model, trained on both corpora.

5 Error analysis

The discrepancy in NLP quality across corpora for different tasks, and especially lower OOD scores below the segmentation level, leads us to suspect that while segmentation in the two corpora is largely mutually consistent (after conversion of the tokenization scheme), other annotation layers may have some issues. Such issues could also explain why joint training does not outperform in-domain scores – improvements from having more training data could be cancelled out by inconsistencies. In order to better understand the most common errors, Table 7 gives the three most commonly confused label pairs for POS tags and dependency relations in both cross-corpus directions (for complete confusion matrices, see the Appendix).

As the table shows, a substantial portion of errors is due to NOUN vs. PROPN confusions, which is not unusual in general (see Behzad and Zeldes 2020) but suspicious given that directions are flipped: testing on HTB we see over-prediction

	train Wiki → test HTB			train HTB → test Wiki		
	gold	pred	freq	gold	pred	freq
<i>POS</i>	NOUN	PROPN	110	PROPN	NOUN	160
	VERB	ADJ	86	VERB	AUX	23
	NOUN	PROPN	73	ADJ	VERB	21
<i>deprel</i>	nmod	obl	77	obl	nmod	73
	compound	nmod	70	nmod	obl	72
	obl	nmod	60	compound	flat	30

Table 7: Top 3 confused tag and deprel pairs by corpus.

of proper nouns, while testing on Wikipedia data shows the opposite trend. Manual inspection of the errors reveals that HTB’s name annotations are inconsistent: names in the original UD HTB data, and especially place names, are tagged NOUN if they are transparently composed of multiple Hebrew nouns. Thus places like *Tel Aviv* (lit. ‘Spring Mound’) or *Kfar Saba* (lit. ‘Grandfather Village’) are not tagged as PROPN in HTB, but are PROPN in IAHLTWIKI, leading to prediction errors in both directions. ADJ vs. VERB confusions are also ‘flipped’, though less common, and are owing to a number of lexical items with different treatments, such as *dome* ‘similar’ and *shone* ‘different’, which are tagged as adjectives in the Wiki data, but as participles in HTB, based on their etymology.

For dependency relations, most confusions are due to obl vs. nmod, which are annotated consistently in the data and correspond to verbal and nominal prepositional modifiers, i.e. the well known challenge of PP attachment (see Kawahara and Kurohashi 2005). Manual inspection confirms that these are less worrying in terms of corpus compatibility. Confusions involving the compound label are mainly due to errors in HTB, which include common items such as *brit ha-mo’atsot* ‘Soviet Union’ (occasionally labeled nmod instead of compound), and often carry an additional annota-

tion `ConvUncertainLabel` in the original data, indicating that this is an issue with the original conversion of the data to UD. We plan to address these and other inconsistencies in future releases of the data.

6 Conclusion

In this paper we make available, describe and evaluate annotation quality in the first non-newswire treebank for Hebrew, complementing the existing newswire data which is by now over 30 years old. We propose and implement a new, simpler tokenization scheme for Hebrew, and release a revised version of the existing UD Hebrew newswire corpus which follows the new scheme. Our evaluation of NLP systems on the new tokenization indicate that it is somewhat easier for a range of tokenizers, at the cost of slightly lower scores for downstream morphosyntactic annotations. We also evaluate agreement on the complete UD annotation pipeline for Hebrew for over 12K doubly annotated tokens, and build a novel infrastructure for error detection based on Grew.

As part of our evaluation of NLP quality on the new resources, we compared several off the shelf systems, whose approaches we combine and incrementally improve to produce a new system, HebPipe, which is released with this paper. Our system achieves new SOTA results on most UD NLP tasks for the older HTB dataset, including segmentation, lemmatization and dependency parsing, evaluated on both the new and older tokenization schemes. We also reported on the first cross-corpus parsing experiment results, which indicated that segmentation and tagging using our system generalize well and substantially better than previous systems, while full dependency parsing is still less reliable out of domain. Error analysis indicates that some of the issues are caused by remaining compatibility problems across the corpora, which we plan to address in the future.

In other future work, we are in the process of annotating data from further sources, including written language from blogs, social media and government websites, as well as spoken data from parliament proceedings and TV shows. To improve annotation, we hope to extend `grewv` to annotator-mediated autocorrection of validation errors including choosing between multiple correction options, as well as machine-guided refinement of annotation guidelines, and guiding source selection for

collecting new material.

URLs for IAHLTWIKI, the revised UD HTB and complete code for the NLP tools from Section 4 will be included after review of this paper.

Limitations

While the resources created for this paper substantially broaden the genres, domains and temporal diversity of data for Hebrew segmentation, tagging and parsing, they are still very limited. In particular, the newly released data reflects relatively formal and well edited language, and does not cover spoken language or informal user-generated content from the web, areas which we would like to explore in future work. Statements about generalizability and OOD reliability of tools should therefore be interpreted cautiously. Our results are only truly reliable for the specific case of the Hebrew datasets examined here, and may not apply to other domains or similar morphologically rich languages (MRLs), such as Arabic, with which we are also experimenting.

The NLP tools compared in this paper also constitute a narrow and subjective selection – the choice of Stanza and Trankit was motivated by their popularity and the existence of previous models for Hebrew, while comparison with [Seker et al.](#)'s work was due to the paper's previous SOTA scores. It is possible that other architectures could outperform the results reported here, as well as reach different conclusions about generalizability and error sources. Due to the large number of pipeline components in the systems compared here, each with separate trained models for each corpus (18 models per experiment across 3 systems), and the focus of this paper on the corpus resource, we decided to only use seeds and not to obtain average scores from a large number of runs, which would require spending substantially more computing resources and promote current carbon-intensive trends in NLP. While we understand the desire for such numbers in papers focused on novel parsing architectures with very close numbers, we feel that they would not change the results presented here (though reviewers can let us know otherwise).

For similar reasons, we performed no hyperparameter optimization on any of the systems, meaning that it is also possible that systems would fare differently if this were attempted, possibly making some of our incremental improvements, such as within-MWT BIES positional embeddings, un-

helpful. That said, we are skeptical that extensive optimization against single-genre dev sets is meaningfully useful for performance on unseen data in the wild, and we prefer to invest in developing systems which can be trained and run with moderate or no GPU resources (see also the Ethics Statement below). Optimizing pipeline architectures thoroughly would mean considering interactions between different components’ models, for gains that would quite possibly be limited to the data used in the paper itself.

Finally, the evaluation of our corpus focused on end-to-end numbers, rather than evaluating each task on gold inputs (i.e. parsing numbers reflect upstream tokenization errors, etc.), but did use gold sentence segmentation, to match the setup in previous SOTA work such as [Seker et al. \(2022\)](#). Especially given the limited space in this paper, we aimed to follow the UD shared task paradigm in preferring end-to-end scores, which more closely mirror expected performance in the wild. Although this complicates the interpretation of better and worse downstream component choices, we feel this is inevitable, since upstream outputs such as segmentation and tagging are also part of the input for downstream tasks, and a large number of evaluation scenarios is conceivable. By contrast, our inter-annotator agreement study required identical tokenization to compute kappa for annotations, meaning the lack of agreement numbers including tokenization disagreements is a further limitation of our results. This last issue likely has low impact on numbers due to the high agreement on tokenization, and we intend to release the doubly annotated raw data for interested researchers as well.

Ethics Statement

This work contributes to open source and open access progress in NLP for morphologically rich languages, and specifically for Hebrew, which has not enjoyed the same wealth of resources as English and other European languages. We recognize that NLP research has a computing cost and carbon footprint, which motivates us to release all models in this work (preventing the need to retrain similar models), to use base-sized language models, and to avoid extensive hyperparameter optimization on these single-genre datasets, which may lead to minor improvements on test sets, but may or may not generalize to applications in the wild.

We also recognize that NLP tools can be used to

do harm, but expect that the type of NLP processing promoted here will do more good than harm by preventing tools from adhering closely to outdated and narrow-domain data, which this work aims to broaden. Given that systems for UD-style outputs for Hebrew already exist, we view any reduction in topical and authorial bias, as well as the public release of more resources, as net positives. All participants in this work have been compensated. The annotators (3 female, 3 male) were employed as regular employees of IAHLT, the non-profit organization which funded the treebank. Previous work has been credited to the best of our knowledge.

Acknowledgements

We are grateful for the kind support of the Israel Innovation Authority and the Israel National Digital Agency. We would also like to acknowledge the immense help of the Israeli Association of Human Language Technologies (iahlt.org), its members and employees for supporting this work, as well as all of the annotators who participated in this project, without whose hard work this resource would not have been possible.

References

- Meni Adler and Michael Elhadad. 2006. An unsupervised morpheme-based HMM for Hebrew morphological disambiguation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pages 665–672, Sydney.
- Giuseppe Attardi, Daniele Sartiano, and Maria Simi. 2021. [Biaffine dependency and semantic graph parsing for Enhanced Universal dependencies](#). In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*, pages 184–188, Online.
- Shabnam Behzad and Amir Zeldes. 2020. A cross-genre ensemble approach to robust Reddit part of speech tagging. In *Proceedings of the 12th Web as Corpus Workshop (WAC-XII)*, pages 50–56, Marseille, France.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

- Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *Proceedings of ICLR 2017*, Toulon, France.
- Yoav Goldberg and Michael Elhadad. 2009. [Hebrew dependency parsing: Initial results](#). In *Proceedings of the 11th International Conference on Parsing Technologies (IWPT'09)*, pages 129–133, Paris, France. Association for Computational Linguistics.
- Gaël Guibon, Marine Courtin, Kim Gerdes, and Bruno Guillaume. 2020. [When collaborative treebank curation meets graph grammars](#). In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 5291–5300, Marseille, France.
- Bruno Guillaume. 2021. Graph matching and graph rewriting: GREW tools for corpus exploration, maintenance and conversion. In *Proceedings of EACL 2021 System Demos*, pages 168–175, online.
- Daisuke Kawahara and Sadao Kurohashi. 2005. Pp-attachment disambiguation boosted by a gigantic volume of unambiguous examples. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP-05)*, pages 188–198, Berlin and Heidelberg. Springer.
- Max Müller-Eberstein, Rob van der Goot, and Barbara Plank. 2021. How universal is genre in Universal Dependencies? In *Proceedings of Treebanks and Linguistic Theories 2021*, pages 69–85, Sofia, Bulgaria.
- Minh Van Nguyen, Viet Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A lightweight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 80–90.
- Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. [Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures](#). In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMC-7) 2019, Cardiff, 22nd July 2019*, pages 9–16, Mannheim. Leibniz-Institut für Deutsche Sprache.
- Siyao Peng and Amir Zeldes. 2018. [All roads lead to UD: Converting Stanford and Penn parses to English Universal Dependencies with multilayer annotations](#). In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 167–177, Santa Fe, NM.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. [Stanza: A Python natural language processing toolkit for many human languages](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108.
- 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.
- Djamé Seddah, Sandra Kübler, and Reut Tsarfaty. 2014. Introducing the SPMRL 2014 shared task on parsing morphologically-rich languages. In *First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages*, pages 103–109, Dublin, Ireland.
- Amit Seker, Elron Bandel, Dan Bareket, Idan Brusilovsky, Refael Greenfeld, and Reut Tsarfaty. 2022. [AlephBERT: Language model pre-training and evaluation from sub-word to sentence level](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 46–56, Dublin, Ireland.
- Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. [Predictive biases in natural language processing models: A conceptual framework and overview](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online. Association for Computational Linguistics.
- Khalil Sima'an, Alon Itai, Yoad Winter, Alon Altman, and Noa Nativ. 2001. Building a tree-bank of Modern Hebrew text. *Traitment Automatique des Langues*, 42:347–380.
- Reut Tsarfaty. 2013. [A unified morpho-syntactic scheme of Stanford dependencies](#). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 578–584, Sofia, Bulgaria. Association for Computational Linguistics.
- Amir Zeldes. 2018. A characterwise windowed approach to Hebrew morphological segmentation. In *Proceedings of the 15th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 101–110, Brussels, Belgium.
- Amir Zeldes and Caroline T. Schroeder. 2016. An NLP pipeline for Coptic. In *Proceedings of the 10th ACL SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH2016)*, pages 146–155, Berlin.
- Amir Zeldes and Dan Simonson. 2016. Different flavors of GUM: Evaluating genre and sentence type effects on multilayer corpus annotation quality. In *Proceedings of the 10th Linguistic Annotation Workshop (LAW X)*, pages 68–78, Berlin.

A Agreement on morphological features

Table 8 details kappa scores for each morphological feature category exhibiting more than 10 disagreements, including both universal UD features, and

Hebrew-specific ones, such HebBinyan (the morphological inflectional class of Hebrew verbs). The IAA study was performed before the introduction of the grewv validation system, and running the system on the double annotated data reveals that in over 29% of the disagreements (387 of 1,297 cases) at least one of the annotators’ decision would have been flagged as an error, based on the rules in Table 9. This result underscores the importance of automatic validation, and the possible lesson that future IAA studies should report the impact of validators on potential disagreement errors that would be prevented by validation.

Label	Kappa	Disagreements
Case	0.902	82
Definite	0.966	123
Foreign	0.434	13
Gender	0.954	281
HebBinyan	0.960	56
Number	0.959	238
NumType	0.688	48
Person	0.966	65
Polarity	0.709	48
Prefix	0.592	11
PronType	0.976	68
Tense	0.968	42
Typo	0.482	15
VerbForm	0.937	37
VerbType	0.703	31
Voice	0.900	139
<i>average</i>	0.818	81.06
<i>total</i>		1297

Table 8: Cohen’s Kappa and the total number of disagreements for each feature and miscellaneous label.

B Technical information and reproducibility

The AlephBERT transformer model by Seker et al. (2022) used for most of the components in this paper is available from huggingface⁸ and was trained over approximately 8 days on a DGX machine (8 V100 GPUs) on close to 18 GB of text from the OSCAR corpus (Ortiz Suárez et al., 2019), Hebrew Wikipedia and Hebrew Twitter. The model uses the standard 12 layer transformer with 768 dimensional word-piece representations or approximately 110 million parameters, with a vocabulary of 52K types.

The flair taggers described above were trained with a biLSTM CRF (n_hidden=256) on top of AlephBERT transformer word embeddings, concatenated with 5 dimensional MWT positional dense embeddings for tagging, and 17 dimensional POS

⁸<https://huggingface.co/models>

Rule name	Feature	Differences
verb-mand-voice	Voice	42
pass-arg	Voice	32
verb-mand-binyan	HebBinyan	28
def-cons	Definite	25
binyan-nomid	Voice,HebBinyan	22
pron-prontype-required	PronType	18
aux-nopolarity	Polarity	17
amod-onlyadj	–	15
aux-binyan	HebBinyan	13
adp-case-gen	Case	13
yesh-polarity	Polarity	12
pron-nopolarity	Polarity	12
noun-adj-gender-agr	Gender	11
nmod-obl-case	Case	11
mandatory-gender	Gender	11
cc-child-conj	–	11
noun-adj-num-agr	Number	10
mandatory-number	Number	10
(other types)	–	74
<i>total</i>		1,297

Table 9: Frequency of validation errors in our inter-annotator agreement study identified by grewv, broken down by type for errors of frequency ≥ 10 . Annotations made after the introduction of our validation system would forbid these disagreements.

embeddings as inputs for morphological features. In all cases we used the default flair hyperparameters with a learning rate of 0.1, optimizing with SGD using a mini-batch size of 15 and halting on dev set accuracy with a patience of 3.

The Diaparser model, also using default hyperparameters, combines fixed AlephBERT embeddings (no fine-tuning) with randomly initialized, fine-tunable fixed embeddings (100 dimensions), which are all fed into a 200-dimensional biLSTM topped by arc and dependency relation MLPs with biaffine attention (500 and 100 dimensional respectively), for a complete parser model with approximately 12M trainable parameters. All training was carried out on a consumer laptop (Dell XPS15, Intel Core i9-9980HK CPU@2.40GHz, 64 GB RAM, NVIDIA GeForce GTX 1650, 4GB GPU RAM).

For Stanza and Trankit’s architectures, as well as RFTokenizer and Seker et al.’s system, we refer readers to the original published papers and tool documentations.

C Label distributions

In compliance with the EMNLP reproducibility checklist, Table 10 gives the exact breakdown of POS tag labels in each dataset.

D Confusion Matrices

The confusion matrices in Figure 2 below show predicted vs. gold label frequencies for POS tagging

	IAHLTWIKI	HTB
ADJ	1,711	1,256
ADP	21,005	21,058
ADV	1,529	1,035
AUX	956	1,014
CCONJ	1,706	1,976
DET	11,177	11,587
INTJ	4	3
NOUN	31,624	31,003
NUM	1,126	1,220
PRON	1,633	1,015
PROPN	11,448	1,181
PUNCT	11,613	11,301
SCONJ	1,317	1,792
SYM	146	1
VERB	11,650	11,359
X	304	118

Table 10: POS label distributions.

and dependency relations in the cross-corpus experiments using this paper’s best system, HebPipe. All numbers are computed using the official CoNLL UD scorer and end-to-end predictions from plain text, taking the scorer’s optimal alignment as the basis for matching predictions and gold token labels, which is non-trivial due to occasional differences in the predicted segmentation of words into tokens.

