The Hebrew Wikipedia Dependency Parsed Corpus Ver. 1.0

Yoav Goldberg

February 6, 2014

This document describes the Hebrew Wikipedia dependency-parsed corpus.

1 The Data

The corpus is based on a September 2013 dump of the sentences in the Hebrew Wikipedia (`http://he.wikipedia.org`), after filtration of tables, headers, and other "non-text" material. The data contains 3,832,992 sentences, and over 76M tokens and over 106M words (106,449,114 tree nodes).

2 Automatic Analysis

The data was tokenized using a simple Hebrew tokenization script by Yoav Goldberg (`https://bitbucket.org/yoavgo/hebtokenizer`) and tagged by the part-of-speech tagger of Meni Adler [1] (the tagger relies on morphological analysis performed by the MILA morphological analyzer [7][1]). The tags were then converted using a deterministic process to the tagset which is used in the Hebrew Treebank (v3). After tagging, the sentences were parsed with the 2013 version of the Hebrew Dependency Parser by Yoav Goldberg. The parser follows the EasyFirst approach ([5, 4]) with error-exploration during training training ([6]).

3 Data Format

All data is in UTF-8 encoding.

Raw The raw text, one sentence per line.

[1] `http://www.mila.cs.technion.ac.il/eng/tools_analysis.html`

Tagged and Parsed Data The tagged and parsed data contains a subset of the sentences in the raw corpus (a length threshold was applied, filtering overly-long sentences).

The parsed sentences are in a format similar to that of the CoNLL-2007 shared task (`http://nextens.uvt.nl/depparse-wiki/DataFormat`). Specifically:

- The data file contains sentences separated by a blank line.
- Each sentence contains one or more words, each word starting on a new line.
- Each word line consists of ten fields, separated by a single tab character. No whitespace is allowed within a field. The fields are described in the table below:

Field	Name	Description
1	ID	Word index, in the form $X.Y$ (see "Word Indexing").
2	WORD	Word form, including pseudo-words (see "Words").
3	LEMMA	Word lemma.
4	CTAG	Coarse-grained part-of-speech tag (see "Part of Speech Tags").
5	FTAG	Finer-grained part-of-speech tag (see "Part of Speech Tags").
6	FEATS	Morphological Features (see "Morphological Features").
7	HEAD	Index of the head-word of the current word.
8	DEPREL	Dependency-relation to the head (see "Syntactic Relations").
9	Reserved	Always _
10	Reserved	Always _

3.1 Word Indexing

The difference from the CoNLL format is that word indices are not integers, but of the form $X.Y$ where both X and Y are integers. This is done in order to support the Hebrew writing system in which some function words are not separated by whitespace but instead attach to the next token. In the indexing scheme used in this corpus, X designates the token number, while Y designates the within-token word number. For example, a sequence such as והילד אמר (and-the-boy said) will be represented as:

1.0 ו
1.1 ה
1.2 ילד
2.0 אמר

3.2 Words

Word boundaries and definitions follow version 3 of the Hebrew Treebank.

Hebrew orthography causes some function words to be attached to the next word as a prefix instead of being separated by whitespace. In addition, the definite article ה is sometimes not explicitly marked. Each function word, including

the definite article, is treated as a separate word in the parse tree (and has its own line / index):

1.0 ו
1.1 מ
1.2 ה
1.3 בית

In addition to the words, some tree nodes are represented as pseudo-words. Pseudo-words are used in cases a word can participate in a syntactic relation, but it does not appear explicitly in the text. For example, the definite marker is not written after the prepositions ב, ל and כ, but is marked using a special diacritic mark on the preposition. In these cases, the definite article appears in the tree using the *DEF* pseudo-word:

1.0 ל
1.1 *DEF*
1.2 בית
2.0 ה
2.1 גדול

Other pseudo-words are used for dealing with morphological suffixes. These include:

Possessive Suffixes ילדותו (his-childhood), which is sometimes analyzed as ילדות של הוא (childhood of him) is analyzed in the corpus as:

1.0	ילדותה	ילדות	NN	NN+POS+PRP	gen=F|num=S
1.1	*POS*	של	POS	POS	_
1.2	*PRP*	היא	PRP	PRP	gen=M|num=S|per=3

Here, *POS* is the possessive, and *PRP* is the pronoun, whose gender, number and person are marked in the morphological features field. Note that the base form and the suffix each have their own set of morphological features.

Inflected Prepositions Words such as להם (to them), לה (to her), בו (in him), בנו (in us) and so on, which are prepositions with a pronomial suffix, are analyzed as the word followed by the pronoun pseudo-word *PRP*:

1.0	להם	ל	IN	IN+PRP	_
1.1	*PRP*	הם	PRP	PRP	gen=M|num=P|per=3

1.0	בנו	ב	IN	IN+PRP	_
1.1	*PRP*	אנחנו	PRP	PRP	gen=M|gen=F|num=P|per=1

1.0	שלכן	של	POS	POS+PRP	_
1.1	*PRP*	אתן	PRP	PRP	gen=F|num=P|per=2

Inflection Accusatives Words such as אותו and אותה (the accusative marker את (AT) followed by a pronomial suffix) are analyzed similarly to the inflected prepositions:

1.0	אותי	את	ACC	ACC+PRP	_
1.1	*PRP*	אני	PRP	PRP	gen=M|gen=F|num=S|per=1

Inflected Adverbs The inflected adverbs עודו and עודה (while he/she) get similar treatment:

1.0	עודו	עוד	RB	RB+PRP	_
1.1	*PRP*	הוא	PRP	PRP	gen=M|num=S|per=3

Verbs with Accusative Suffix In verbs such as אבקשכם (I will ask you) in which the object is realized as a suffix, both the accusative marker and the pronoun are realized as pseudo-words.

1.0	אבקשכם	אבקש	VB	VB+ACC+PRP	gen=M|gen=F|num=S|per=1
1.1	*ACC*	את	ACC	ACC	_
1.2	*PRP*	אתם	PRP	PRP	gen=M|num=P|per=2

1.0	לקבלו	לקבל	VB	VB-TOINFINITIVE+ACC+PRP	_
1.1	*ACC*	את	ACC	ACC	_
1.2	*PRP*	הוא	PRP	PRP	gen=M|num=S|per=3

Summary In all these cases, the suffix is not separated from its base word, and instead the complete form is followed by the pseudo-words of its suffix. The benefit of the pseudo-words system is that, after ignoring the pseudo-word tokens, the word sequence has one-to-one correspondence to the raw tokens.

3.3 Part of Speech Tags

The fine-grained parts-of-speech tags are listed in Table 1. The coarse-grained parts-of-speech column include a coarser version of the fine-grained tag. For words that include suffixes, the fine-grained tag encodes information about the parts of speech of the base as well as the suffix, separated by the + sign.

We note that the tags are lexical, and not syntactic.

3.4 Morphological Information / Features

The sixth field (FEATS) contain morphological and other features of the word. The this field is a bar ("|") separated list of key-value pairs of the form `key=val`, or "_" in case of an empty list.

Features with different possible values (for example, in case the gender can be both male or female) are represented as multiple occurrences of the same key with different values.

The possible keys and values are summarized in Table 2.

Tag	Description	Examples
ACC	AT accusative marker	את, *ACC*
ADVERB	Adverb appearing as prefix	כ
BN	Beinoni (participle) form	מדובר, נוגע, נוסף, קשור
BNT	Construct-state Beinoni Form	מוכי, מנחילי, אוזלת, מרבה, מטבע
CC	Conjunction	כפי, ככל, בניגוד, ש
CC-COORD	Coordinating Conjunction other than ו	רק, אבל, או, אם, גם
CC-REL	Relativizing Conjunction	אשר
CC-SUB	Subordinating Conjunction	כמו, כאשר, לאחר, כדי, כי
CD	Number	אחת, 1, 0
CDT	Construct Numeral	עשרות, מאות, אלפי, שתי, שני
CONJ	The ו coordinating word	ו
COP	Copula (present) and Auxiliaries (past and future)	יהיה, אינו, היתה, היו, היה
COP-TOINFINITIVE	Infinitive Copula	להיות
DEF	H marker	ה, *DEF*
DT	Determiner (other than ה)	כל, מבחר, איזושהי
DTT	Construct-state Determiner	הרבה, שום, אותו, כמה
EX	Existential	היה, ישנם, אין, יש
FW	Foreign Word	England, wiki, New
IN	Prepositions	בין, מתוך, על, כגון
INTJ	Interjection	אוי, נא, חלילה, אוף, פוס
JJ	Adjective	לאומי, גדול, חדש, רבים, אחרים
JJT	Construct-state Adjective	מרובי, חסרת, רעולי, מודעי, מדובר
MD	Modal	עלול, יכולה, צריך, יכול, אפשר
NCD	Numerical Expression	20.30, 20.00, 16.11, 6.6.82, 13.11
NEG	Negation	אין, לא
NN	Noun	ארץ, יום, ממשלה, משטרה
NNP	Proper Nouns	אביב, כהנא, ירושלים, ישראל
NNT	Construct-state nouns	ידי, וודעת
POS	Possessive	של, *POS*
P	“Prefix” wordlets	אנטי, בין, תת, אי, בלתי
PREPOSITION	Suffix (משכל”ב) Prepositions	כש, כ, מ, ל, ב
PRP	Pronouns	זו, הם, היא, זה, הוא, *PRP*
QW	Question Word	היכן, מ, האם, מי, מה
RB	Adverbs	אתמול, כבר, עוד, יותר, לא
REL-SUBCONJ	Relativizer	ש
TEMP-SUBCONJ	Temporal Suboordinating Conjunction	מש, כש
VB-TOINFINITIVE	Infinitive Verbs	לבצע, לתת, למנוע, לשלם, לעשות
VB	Verbs	יודע, נראה, אומר, אמר
yyCLN	Colon	:
yyCM	Comma	,
yyDASH	Dash	–
yyDOT	Period	.
yyELPS	Ellipsis	…
yyEXCL	Exclamation Point	!
yyLRB	Open Brackets	(
yyQM	Question Mark	?
yyQUOT	Quotation Mark	“
yyRRB	Close Brackets	)
yySCLN	Semicolon	;
UNK	Tagger uncertainty	

Table 1: Fine-grained Part-of-Speech Tags

Feature	Key	Possible Values
Gender	gen	`M` (male), `F` (female)
Number	num	`S` (single), `P` (plural), `D` (dual)
Person	per	`1`, `2`, `3`, `A` (all)
Tense of verbs	tense	`PAST`, `BEINONI` (present), `FUTURE`, `IMPERATIVE`
Polarity	polar	`neg` (negative), `pos` (positive)
Type of pronoun	type	`DEM` (demonstrative), `PERS` (personal), `IMP` (impersonal)

Table 2: Morphological and other Features

3.5 Choice of Heads

The choice of heads follows the decisions in Chapter 5 of [3].

3.6 Syntactic Relations

The syntactic relations follow the scheme put forward by Reut Tsarfaty in [8], which is an elaboration of the Stanford-dependency labels [2]. Note, however, that while the relations are close to the Stanford-Dependencies scheme, the choice of heads sometimes differ (this may change in future releases).

The list of relations is detailed below. Relations marked with a * are not very frequent, and are of less important. Some relations are accompanied by examples. The example is given in the format *rel(head,modifier)*.

advmod Adverbial modifier.

... המתמטיקה מתפתחת במקביל כענף
advmod(מתפתחת, במקביל)

amod Adjectivial modifier.

יישומים מעשיים
amod(יישומים, מעשיים)

appos Apposition.

... חבר הוועדה, פול, שדחה
appos(פול,חבר)

also used for parenthesized expressions (which are headed by the opening paren):
בת דודתו (משני הצדדים)
appos(בת, ( ))

aux Auxiliary verbs.

הוא היה זוכה במדליה
aux(זוכה, היה)

... ניתן היה להעניק
aux(ניתן, היה)

cc Introducing conjunction – relation between a coordinating conjunction to its head (when there is no better alternative, usually under root).

conj Relation between the conjoined element and the coordinator.

וועדת ה עבודה ו ה רווחה
conj(ו, עבודה)
conj(ו, רווחה)

ccomp Complement clause with internal subject.

היא הודיעה כי הוועדה תגבש הצעת חוק
comp(הודיעה, כי)
ccomp(כי, תגבש)

comp Complement.

היא הודיעה כי הוועדה תגבש הצעת חוק
comp(הודיעה, כי)
ccomp(כי, תגבש)

הדברים מעידים על הכותב
comp(מעידים, על)
pobj(על, הכותב)

complmn * Used for a complementizer which is the root of the sentence.

compound * Used in compositional adverbs such as לטובת, לרוחב, במהירות in cases they receive a split analysis.

אסיפת בחירות לטובת מועמד
prepmod(אסיפת, ל)
compound(ל, טובת)
pobj(טובת, מועמד)

cop Copular element – modifying the head of the copular expression.

ישראל היא מקום מעניין
cop(מקום, היא) subj(מקום, ישראל)

def Relation between definite article and its head.

dep A generic dependant (governed, or an argument).

det Relation between a determiner (other than the definite article) and its head.

כל ה ילדים
def(ילדים, ה)
det(ילדים, כל)

detmod * Modifier of a determiner or of a quantifier.

gen Genitive.

ילדים של קיץ
gen(ילדים, של)

ghd * Genitive head of a nominal phrase.

gobj Genitive Object, either of a possessor or in construct-state (סמיכות).

השלבים הראשונים של לימודי היהדות
gobj(לימודי,של)
gobj(יהדות,לימודי)

hd A generic head (governor, argument taking).

mod A generic modifier. Super-type of appos, amod, advmod, neg, prepmod, possmod, tmod, rcmod, infmod

mwe * Part of a multiword expression (out of a relatively small set of recognized multiword expressions).

neg Negative modifier.

לא נותר לו זמן
neg(נותר, לא)

nn Noun-noun modification, mostly within proper names.

יוסי שריד
nn(יוסי, שריד)

ארצות ה ברית
nn(ארצות, ברית)

null *

num A modifier which is a numeric expression.

שש ה שנים
num(שנים, שש)

number Part of a number.

obj Genitive object (when dobj, iobj, gobj, pobj are not appropriate or cannot be determined). Usually object of an infinitive verb.

לאסוף כסף
obj(לאסוף, כסף)

parataxis

pcomp * Complement clause of a preposition.

נאבקים כדי לזכות בהכרה
comp(נאבקים, כדי)
pcomp(כדי, לזכות)

pobj Object of a preposition.

הוא הלך ל ביתו
pobj(ל, ביתו)

posspmod * Possession modifier.

prd Verbal predicate (usually at root of sentence).

prep * A preposition (which is not in head position, unlike prepmod and comp). Usually used for prepositions under root.

prepmod A preposition modifier. הוא הלך ל ביתו prepmod(הלך, ל)

punct Relation between a punctuation symbol and its head.

qaux * Question auxiliary. Usually question-words at root.

rcmod Introducing a relative-clause modification. הילד ש לבש חולצה צהובה

rcmod(ילד, ש) relcomp(ש, לבש)

rel *

relcomp Complement of the relativizer. הילד ש לבש חולצה צהובה

rcmod(ילד, ש) relcomp(ש, לבש)

subj Subject. הוא הלך לשם אתמול

subj(הלך, הוא)

tmod * Temporal modifier. העד הושבע אתמול

tmod(הושבע, אתמוך)

xcomp Complement clause with external subject (usually in infinitive constructions). המכונה נועדה לגרום נזק רב

xcomp(נועדה, לגרום)

References

[1] Meni Adler. *Hebrew Morphological Disambiguation: An Unsupervised Stochastic Word-based Approach*. PhD thesis, Ben-Gurion University of the Negev, 2007.

[2] Marie-Catherine de Marneffe and Christopher D. Manning. Stanford dependencies manual. Technical report, Stanford University, 2008.

[3] Yoav Goldberg. *Automatic Syntactic Processing of Modern Hebrew*. PhD thesis, Ben-Gurion University of the Negev, 2011.

[4] Yoav Goldberg and Michael Elhadad. Easy-first dependency parsing of modern Hebrew. In *Proc. of NAACL Workshop on Statistical Parsing of Morphologically Rich Languages*, 2010.

[5] Yoav Goldberg and Michael Elhadad. An efficient algorithm for easy-first non-directional dependency parsing. In *Proc. of NAACL*, 2010.

[6] Yoav Goldberg and Joakim Nivre. Training deterministic parsers with non-deterministic oracles. *Transactions of the association for Computational Linguistics*, 1, 2013.

[7] Alon Itai and Shuly Wintner. Language resources for Hebrew. *Language Resources and Evaluation*, 42(1), March 2008.

[8] Reut Tsarfaty. A unified morpho-syntactic scheme of stanford dependencies. In *Proc. of ACL*, 2013.