Hebrew Treebank 1.0

Background

The Hebrew Treebank contains 4,783 sentences (88,747 words, 120,213 morphemes) of news items from the Ha'aretz daily newspaper, with full segmentation into morphemes and morpho-syntactic analysis. Morphological features that are not directly relevant for syntactic structures, like roots, templates and patterns, are not analyzed. A version of the Treebank with only the morphological level is also supplied.

The Hebrew Treebank was designed with a POS tag set and a syntactic tag set that are as close as possible to those of the **English Penn Treebank**. A significant difference from English is that in the Hebrew Treebank the annotated words are separated into morphemes, which may have different POS tags and syntactic positions in the tree.

For example, the two words BBIT HGDWL ("in the big house") are analyzed as five different morphemes:

B	H	BIT	H	GDWL
in	the	house	the	big

Note that the first occurrence of the morpheme H ("the") is covert in the word BBIT. Segmentation into morphemes makes it possible to analyze different morphemes of the same word as belonging to different constituents in the tree:

[B [[H BIT] [H GDWL]]

Another significant difference from English is the relatively free order of constituents in Hebrew. In order to encode syntactic functions of constituents within a tree structure, functional features (for subject, object etc.) were added to constituents, and the constituents themselves appear in a "flat" order.

Some tags were added to the tag set of the Penn Treebank, to accommodate special properties of Hebrew like construct states, accusative marking (AT), the definiteness morpheme H, and verbless predicates.

For a detailed description of the linguistic annotation scheme see Sima'an et al (2001): http://mila.cs.technion.ac.il/treebank/tal.pdf

For more examples and conventions see the annotator guidelines.

Software

The following software was used for the development of the treebank:

SEMTAGS (Remko Bonnema, University of Amsterdam)
A GUI for tree annotation.

Morphological analyzer (Erel Segal, Technion)
Provides an initial tagging for the input sentences, which was manually fixed by the annotators.

Format translation script (Alon Altman, Technion)
Translates the output of the morhpological analyzer to the Treebankλ~@~Ys format, and creates initial flat trees for annotation.

Mapping trees to the original text (Avihai Dgani, Technion)
This program maps each word in the raw text to the corresponding morphemes in the treebank, and checks the synchronization between the parse trees and the original text.

Comments

Null trees: the original text was automatically segmented into sentences. In some cases, a single sentence in the original text was splitted into multiple sentences. Therefore, in the annotation process it was necessary to rejoin sentences. In order to maintain the synchronization between the tree numbers and the automatically segmented sentence numbers, null trees were inserted. For example, if the first sentences was splitted into three sentences, #1, #2 and #3, then in the treebank, tree #1 will include the three sentence parts, while tree #2 and tree #3 will be null trees.

A null tree looks like this:

S
|
yyDOT
|
yyDOT

and is represented as: ((S (yyDOT yyDOT)))

Duplicate sentences: sentences 24-36 in the original text do not appear in the treebank, since they are repeated in sentences 1358-1370. Sentences 1249-1293 do not appear in the treebank, since they are repeated in sentences 1204-1248.

Missing trees: the following sentences currently do not appear in the treebank (represented by null trees): 552, 772, 1350, 1382, 2044, 3206

Transliteration

א	ב	ג	ד	ה	ו	ז	ח	ט	י	ך	כ	ל	ם	מ	ן	נ	ס	ע	ף	פ	ץ	צ	ק	ר	ש	ת	"	%
A	B	G	D	H	W	Z	X	J	I	K	K	L	M	M	N	N	S	E	P	P	C	C	Q	R	F	T	U	O

Staff

Research supervision

Prof. Alon Itai, Prof. Yoad Winter, Dr. Khalil Sima'an

Development of the annotation scheme, and tree annotation:

Yair Adiel, Nomi Guthmann, Shiri Kenan, Adi Milea, Noa Nativ, Roni Tenzman, Pnina Veisberg.

Technical Support

Alon Altman, Roy Bar-Haim, **Shlomo Yona**.