Hebrew Morphological Tagging Guidelines

BGU Computational Linguistics Group

December 2008
Beer-Sheva

Revision number: 001

Contents

Chapter 1

Introduction

This document provides guidelines to people who tag part-of-speech (POS) data on Modern Hebrew text. The objective of the manual tagging is not to provide full morphological information on the words in the text, but to provide the minimal information so that the word can be analyzed without ambiguities.

The objective is that human taggers who follow these guidelines will reach high agreement when tagging all the words in arbitrary text in written Modern Hebrew. The document presents the specific Tagset we have designed for Modern Hebrew, and for each tag, it explains which words are to be tagged, and how to avoid confusing the tag with other closely related tags.

1.1 Tagging Scenarios

The guidelines are designed to help people tag in the following two scenarios. In both cases, we assume that tokens and sentences have been identified before the tagging procedure starts.

1.1.1 Manual Tagging

A source text is provided in a file; it is separated in sentences and tokens (cf. below for the definition of tokens). Taggers use a manual tagging utility to annotate the text (we currently use WordFreak, available at `http://wordfreak.sourceforge.net/`). A tag is associated to every token in the text. The tagging result is encoded as an XML file which can then be further analyzed by automatic tools.

1.1.2 Online Tagging

In online tagging, the tagger works interactively with an automatic morphological analyzer. The morphological analyzer iterates through a source text where tokens and sentences are already marked up. For each token, the analyzer returns the list of all possible morphological analyses. The analyses are displayed to the manual tagger according to the schema described in this document. Internally, there is more information, but the information is displayed in as 'compact' a format as possible to avoid 'information overload'. The tagger selects the analysis that he finds appropriate according to the context. In rare cases, the tagger enters an analysis that the morphological analyzer did not propose.

We are currently integrating the Hebrew morphological analyzer of MILA [13] with WordFreak to support this mode of online tagging.

1.1.3 Precision vs. Detail Level

Manual tagging adds value to an annotated corpus. This value comes in two forms:

- Detailed information: human taggers can distinguish among close variants, and provide rich details about linguistic data.
- Reliable data: human taggers provide the reference judgment agains which automatic procedures will be judged in the future, and from which automatic learning procedures will proceed.

Between these two objectives, we give preference to reliability over detail. This preference means that when we decide on a guideline - if the guidelines to decide how to tag a certain construct or how to distinguish between two tags becomes too complex, we must revise the tagset and prefer a tagset where complex decisions are avoided.

During the development of these guidelines, we checked that the guidelines provide answers to taggers so that taggers eventually agree on their tagging in a highly consistent manner.

1.2 The Tagset

The notion of *part of speech* has a long tradition.[1]. The central parts of speech (verb, noun, and adjective) are part of the basic linguistic intuitions of all speakers.

[1]For a historical review see [12, *p.* 203–209]; for a discussion on the English parts of speech set, see [31, chapter 2].

Surprisingly, however, a complete list of parts of speech for a given language is not well established. Beyond verb, noun, and adjectives, many other lexical units appear in text – and each raises potential questions as to what is meant by part of speech, what is the best way to label every unit in a document, and how to distinguish among the various labels.

In English, for example, various lists of parts of speech have been used in various tagging projects, where the size of these tagsets range from 48 to 195. The Penn Treebank [21] reduces the 87 tags of Brown corpus [14] to 36 POS tags and 12 other tags. The reduced set leaves out information that can be recovered from the identity of the lexical item. Most tagging situations, however, do not involve parsed corpora and require a larger tagset. Corpora that aim to code more grammatical behavior use a much larger tagset, such as the Lancaster-Oslo/Bergen corpus (135 tags), the Lancaster UCREL group (about 165 tags), and the London-Lund Corpus of Spoken English (197 tags). The relation between these and some other tagsets is discussed in [15, Appendix B].

According to Ornan [25], the part of speech is an integral attribute of the morphology. Parts of speech were designed in order to categorize the type of the words. Ornan argues against the involvement of syntactic consideration in the definition of standard parts of speech sets in various languages. [2] Schwarzwald [27, volume 1, *p.* 100–109][3] presents the traditional list of nine parts of speech for Hebrew.[4], which is equivalent to the classic list of Gesenius [16]:

- Nouns: common nouns, adjectives, pronouns, numerals.
- Verbs
- Particles: prepositions, conjunctions, interjections, adverbs.

As noted by Schwarzwald, the particles cover 'closed sets' of words (*i.e.*, those which undergo few changes over time), where as nouns and verbs are open sets.

Rosen [26, chapter 5] defines four categorical dimensions – *person-sex-number*,[5] *gender-quantity*, *case*,[6] and *tense* - in order to formalize thirteen word classes (*i.e.*, parts of speech), as shown in Table 1.1. The finite verb category, for instance, is

[2] For a discussion on POS categorization criteria, see [19, chapter 2].

[3] See also [6, chapter 9].

[4] For a description of this categorization in term of Goshen's theory, see [27, volume 1, *pp.* 151–153].

[5] See below.

[6] *i.e.*, prefixation of ב, ל, את, על, etc. – see [26, section 7.1.4].

Part of speech		person-sex-number	gender-quantity	case	tense
Verb	**Finite**	X			X
	Infinitive	X			
	Gerund	X		X	X
	Participle		X		X
Noun	**Appellative**	X	X	X	
	Personal pronoun Anthroponymic	X		X	
	Toponymic		X	X	
Adverb	**Local-temporal**			X	
	Other				
Verboid		X	X		X
Impersonal					X
Adjective			X		

Table 1.1: Word categorization according to Rosen's four categorial dimensions.

composed of all forms, inflected by person/sex/number/tense, in contrast to participles which have no person dimension (but do have tense mark, *e.g.*, היורדים הוחזרו). Gerunds differ from infinitives due to their syntactic tense mark (בלכתו קמה), and the attachment to case prefixes (צפיתי ברדת הטל). Nouns can attach case morphemes, and have no tense dimension. **Appellatives** have gender/quantity dimension,[7] and can be inflected by person/sex/number of the possessive pronoun suffix.[8] **Personal pronouns** and **anthroponymics** are not inflected by gender/quantity, where **toponimycs** are not suffixed by possessive pronoun with person/sex/number. **Adjectives** are inflected by gender/quantity, with no case attachment. **Impersonals** have no inflection but tense,[9] where adverbs can be subcategorized to **local-temporals** which attach case morphemes, and all **other adverbs** that are not inflected at all. Finally, **verboids** have all dimensions but case, *e.g.*, היה לי, היה לך, היה להם, היו לנו, יהיו לכם, יש לה, יש לו, etc.

The main contribution of Rosen's elegant work is the positioning of the morpho-syntactic properties as the base criteria for lexical classification. Some of these criteria can be used for formalizing tests for resolving the part of speech of a word in a given context.

In the scope of building a morphological analyzer, the following points should be taken into account:

[7] *e.g.*, ילד, ילדה, ילדים, ילדות.

[8] *e.g.*, ילדי, ילדיך, ילדן.

[9] *e.g.*, צריך לישון -- היה צריך לישון.

1. Even though the dimensions suggested by Rosen are basically morphological, the implementation for lexical classification, is mostly syntactic. The tense mark of impersonals, *e.g.,* צריך, for instance, is given by the syntactic structure of the whole phrase.

2. Some of Rosen's definitions are not clear. The distinction between toponymics and anthroponymics seems to be unnatural. Why are anthroponymics considered to have person dimension and toponymics not? How do temporal-local adverbs attach a case morpheme?

3. An attempt to apply his method for the task of tagging a corpus exposed some deficiencies. We found many adjectives that do attach case, *e.g.*, לעייפים. Rosen may argue for a hidden noun, *i.e.*, לאנשים) העייפים, but such a policy would make the analysis complicated. The lexical category of numerals – some are inflected by person/sex/number[10] and some are not[11] – is not clear. There is no part of speech for punctuations, interjections, date and time, foreign words, titulars, and URL addresses.

4. Rosen does not classify tokens but words, which are sometimes composed of morphemes of several tokens, *e.g.*, היה לו, יהיה צריך. This could be problematic for traditional morphological analyzers, which are based on tokens.[12]

Several computational analyzers were developed in the past decade, such as [8], [28],[13] [34],[14] [9],[15] and [30],[16]. The parts of speech set for each of these analyzers are summarized in Table 1.2.

One of the objectives of this document is to define an exhaustive list of tags - that is, for any document in Hebrew, we must be able to assign non-ambiguously a single tag to every word in the document. This includes typos, proper names, acronyms, foreign words, interjections, abbreviations, numbers, units and all the irregular little words found in documents.

Before we present the tagset we have designed, we must define what are the units to be tagged (and therefore distinguish between words and tokens), and what are

[10] שנים - שניהם

[11] מאה

[12] See See [1, section 6.3.1] for a text representation which can flexibly handle such inter-token words.

[13] `http://www.cs.technion.ac.il/~erelsgl/bxi/hmntx/teud.html`

[14] `http://cl.haifa.ac.il/projects/hebmorph`

[15] `http://www.ravmilim.co.il`

[16] `http://mila.cs.technion.ac.il/website/english/resources/corpora/treebank`

Part of Speech	HMA	Segal	Yona	Rav Milim	Treebank
Noun	X	X	X	X	X
Pronoun	X	X	X	X[a]	X[b]
Proper Name	X	X	X	X	X
Adjective	X	X	X	X	X
Verb	X	X	X	X	X
Adverb	X	X	X	X	X[c]
Preposition	X	X	X	X	X[d]
Conjunction	X	X	X	X	X
Numeral	X			X[e]	
Quantifier		X	X		X
Determiner					X
Aux Verb		X			X
Interrogative	X	X		X	X[f]
Interjection			X	X	X
Particle	X	X			
Prefix				X	X
Suffix				X	
Negation			X		X
Abbreviation	X				
Punctuation	X		X		X
Foreign	X				X
Existantial					X
Modifier[g]					X
Explanation				X	

[a] independent,indefinitive,demonstrative,interrogative

[b] With two syntactic notions: PRP,AGR - see: `http://mila.cs.technion.ac.il/treebank/Decisions-Corpus1-5001.v1.0.doc`

[c] With two syntactic notions: RB,RBR - *Ibid.*

[d] With three syntactic notions: IN,POS,AT - *Ibid.*

[e] cardinal,ordinal,distributive

[f] With three syntactic notions: QA,HAM,WDT - *Ibid.*

[g] Used for general modifiers of nouns, adjectives, adverbs, prepositional phrases - *Ibid.*

Table 1.2: Parts of speech sets of various computational analyzers for Hebrew.

the criteria we use to distinguish among different tags (that is, define what we mean by parts of speech).

1.3 Words vs. Tokens

Hebrew morphology allows morphemes to be combined systematically into complex word-forms (for an overview of Hebrew word forms see [1, Sections 2.2-2.4]). In different contexts, the same morpheme can appear as a separate word-form, while in others it appears agglutinated as a suffix to another word-form ברכו – ברך אותו, שולחנו – שולחן שלו.

Another class of morphemes always appear agglutinated as prefixes to word-forms (for example, prepositions). But, not all instances of the same word-class appear agglutinated (some prepositions, *e.g.*, ב-, מ- are agglutinated, others are not על).

A final complication is due to the fact that when morphemes are prefixed or suffixed to a word-form, the resulting word-form can become ambiguous between an analysis where the suffix is an agglutinated morpheme or a fusioned morpheme ברק – ב-רק, ברכ-ו(pronoun accusative suffix), ברכו (past 3rd person suffix inflection).

The task of the part-of-speech analyzer is to identify each morpheme in a text and tag it. We distinguish between two types of morphemes:

- Morphemes which correspond to word-classes that can appear as distinct word-forms (known as *free morphemes* - in Hebrew, prepositions, accusative pronouns, possessive pronouns)

- Morphemes which correspond to fusion contexts only and cannot appear as separate word-forms (known as *bound morphemes* - in Hebrew, person, number, tense, gender indicators in verb inflections, number, gender, construct state indicators for nouns, number and gender for adjectives).

Accordingly, we expect the tagger to segment each word-form (which we call a 'token') into a sequence of morphemes of the first type prefixed and suffixed to a base form: ל-הודיע-ם → להודיעם

We refer to the original word-form as a token, and the resulting units as 'words'. Each word is characterized by a base-form, a part of speech and inflection features.

The human tagger must provide sufficient information to recover non-ambiguously the segmentation of a given word-form. We found that it is sufficient in general to indicate how many letters belong to the prefix sequence and whether a suffix is present to recover completely the analysis of prefixes and suffixes. Consequently, we only

require the human tagger to indicate the number of prefix letters if they are found and a boolean flag when a suffix is found: (p1, suf) → ל-הודיע-ם → להודיעם

1.4 Tagging Prefixes

The tagger does not need to specify all the analysis of the prefix and of the suffix - which can be recovered automatically except in the following cases:

The prefix ה can be tagged in 3 different forms:

- Definite article
- Relativizer
- Interrogative

In addition, in some cases, the definite article does not appear explicitly: ב ה בית → בבית (in the house)

In all cases where the definite article is analyzed - either explicit or implicit - we require the tagging of a prefix form (H).

The interrogative form of the prefix ה is tagged with a different tag (pint).

The prefix כ can be interpreted in two ways:

1. As a preposition (used in comparative constructs)
2. As an adverb (meaning *approximately*).

We require the comparative usage to be tagged (with a K tag).

In the following cases, the word-form consists only of a prefix (preposition) and a suffix (pronoun) - that is, there is no base component in the word form: בו, עליו. By convention, we tag these combinations as base+suffix (as opposed to prefix+base which would also be applicable).

1.5 Tagging the Lemma

In some cases, word forms can appear with the same form but correspond to different lemmas. For example, שבתה can be analyzed as 'her shabat' or 'her sitting'.

To distinguish among these, when working in the 'online tagging' scenario, the lemma is part of the proposed tag. When working in the 'manual tagging' scenario, we do not require the tagger to enter the lemma.

1.6 Continuation Words

In the same way as agglutination causes difficulty when tagging, compound words are also difficult to tag consistently. A compound word-form contains several word-forms, which appear separated by the usual orthographic delimiters, but has a separate meaning or different properties than expected by the usual combination of its units. For example, the words אף על פי ,על יד are single word-forms made up of words that could appear independently (אף ,על etc).

Several strategies can be used to tag such compounds:

- Tag each sub-word independently of the fact that it appears in a compound. For example, אף would be tagged in the context of אף על פי in the same way as it would be tagged in a context where אף would appear independently.
- Tag the whole sequence as a single unit.
- Tag each word-form in the sequence with a special marker indicating that the word-forms belong to a larger unit.

We opted for the 3rd option because of difficulties found in the first two options:

1. In the example of אף - the word in isolation is ambiguous between a noun (nose) and an adverb (even). None of these two tags are appropriate to the tag of the larger unit (conjunction). Human taggers consistently refused to tag the smaller units in this way. Another problem with this approach is that it loses information - the fact that a word-form belongs to a larger unit is not easily recoverable in most contexts.
2. In order to tag the whole sequence as a single unit, one must allow the tagger to define the unit boundaries. This was practically difficult and proved counter-intuitive to human taggers.

The specific tagging method we use is to tag the first word of the sequence with the tag of the whole sequence, plus a specific tag *cont* (for *continued*). The next elements in the sequence are simply tagged as *cont*. This usage corresponds to the *ditto* tag used in the LOB corpus [20, *p.* 130].

The boundary between a legitimate compound expression and regular compositional construct is notoriously slippery. We, therefore, provide strict guidelines to avoid disagreements:

- Names of location are tagged as continued: תל אביב תל-אביב

- Names of persons are tagged as continued when the first name or family name is compound - else they are marked as two distinct proper nouns: יצחק רבין *vs.* בן אליעזר.
- Construct state (smichut) is never tagged as continued - even for very strongly established expressions בית ספר, עורך דין.
- A closed list of compound prepositions and conjunctions is provided below (in the section on conjunctions and prepositions).

1.7 Decision Criteria

When tagging all units in a document, there are a few word classes word-forms for which taggers consistently feel the task of selecting a tag is difficult. These guidelines focus on the 'difficult' cases by providing tests to decide how to select the most appropriate tag. In general, a tagging difficulty is caused by a possible 'confusion' between close tags.

The tests we provide to distinguish between the confusion pairs involve information at several level:

- Semantic: word-classes often have a similar semantic function - for example, they denote the same type of object or action, modify predicates of different types. When a word-form is difficult to categorize, we may decide which tag is most appropriate by comparing its semantic function in the context to the one of a clear-cut case.
- Syntactic: words in the same class fulfill similar syntactic functions within a sentence. To decide which tag is most appropriate, we may replace the difficult word-form with another word fulfilling the same syntactic function. Another syntactic test consists of conjoining the word with another word of the same syntactic category.
- Morphological: words belonging to the same class behave in a similar manner in terms of morphological agreement and inflection. To decide on the tag, we may replace it by another word whose inflection provides clear-cut indication.
- Applicative: the decision to use a certain tag may help a stochastic tagger to reach a conclusion on the words neighboring it. For example, assuming we use a tag for 'determiner' - this will likely impact strongly on the tagger decision to tag a subsequent word as a 'noun'. For other applications, other criteria

could be used to determine whether a distinction is recoverable and useful. For example, in the context of an information retrieval (IR) application, [8] define which criteria are relevant to help improve IR performance.

As a general rule, when designing a tagset, introducing 'fine-grained' tags makes the tagging process more difficult - because finer decisions must be taken - but as a tradeoff, some fine-grained tags may help the overall process by providing information on the context.

For example, consider how the word יש should be tagged. We found that the word-form יש can be used in three different contexts: existential/locative, modal and possessive. (These contexts differ on the semantic and syntactic dimensions.) Should the tagset include a general יש tag for all occurrences of the word-form, or should we split it into the three context classes we identified? We decided to tag יש according to the context usage (as one of the three options listed above). This distinction means we are introducing 'fine-grained' distinction for the same word-form. This decision is justified by the expectation that the different tags correspond to different syntactic contexts and can be easily identified by a stochastic tagger.

When we had to make such decisions (e.g., whether to split the יש tag into 3 distinct tags), we identified a large set of occurrences of such contexts from a test corpus and verified that human taggers agreed on the tagging according to the guidelines we provided.

1.8 Methodology

While converging on the tagset presented in this document, we iterated through the following process:

1. Establish a base Tagset.
2. Tag manually a set of documents - several taggers for each document.
3. Review unknown or disagreement cases.
4. Establish decision tests.

This document presents a snapshot of the result of this process after approximately 20 iterations with a group of 8 taggers. At each iteration, 20 to 50 documents of between 100 and 500 words were tagged by the group.

Chapter 2

List of Tags

2.1 Parallel Tags

p1, **p2**, **p3**, **p4**, **p5** – number of prefix letters: pn = n letters in the prefix.
h – hidden or explicit definite article ה.
pint – interrogative ה.
suf – suffix.
cons – construct.
cont – continued tag.
k – comparative prefix.

2.2 Nouns

NN – noun gender unspecified singular.
NNS – noun gender unspecified plural.
NNF – noun feminine singular.
NNM – noun masculine singular.
NNFS – noun feminine plural.
NNMS – noun masculine plural.
PNM – proper noun.

2.3 Adjectives

JJF – adjective feminine, singular.
JJM – adjective masculine, singular.
JJFS – adjective feminine, plural.
JJMS – adjective masculine, plural.

2.4 Verbs

VB – verb in infinitive.
VBB – verb in bare infinitive form (makor).
VBP – verb in past tense.
VBR – verb in present tense.
VBF – verb in future tense, masculine form.
VBF-f – verb in future tense, feminine form.
VBI – verb in imperative.

2.5 Others

BN – beinoni (participle).
RB – adverb.
DT – determiner.
– number.
EX – existential.
CP – copula.
NEG – negation.
MD – modal.
INT – interrogative.
UH – interjection.
CONJ – conjunction.
PREP – preposition.
PRO – pronoun.
PREX – prefix word.
FW – foreign word (not proper nouns).
PUN – punctuation.

JUNK – unrecognizable token.
URL – url.

Chapter 3

Definitions and Examples

3.1 Parallel Tags

We define a token as a string of letters separated in writing by punctuations or spaces from other tokens. A token is annotated as a base form tag plus potentially several parallel tags, which denote the prefixes, suffixes, construct state and whether the token is part of a larger unit.

3.1.1 Prefix letters and h

Each letter counts as a single prefix. **h** is added to words in which the definite article ה is implicit or explicit.

(3.1)
כשהבית [p3+h][1]
הלכנו לבית [p1+h] הגדול [p1+h]

3.1.2 Interrogative prefix - pint

The interrogative form of ה is tagged with **pint**

(3.2)
השומר [p1+pint+const+nnm] אחי אנוכי?
הרצחת [p1+pint+vbp] וגם ירשת?

[1] כש is a single prefix but we count letters

3.1.3 Comparative prefix - k

When the prefix כ is used in the sense of a comparative preposition, it is tagged with the K tag.
Note that כש- is sometimes used as the sequence of כ and ש, and sometimes as a single prefix כש (for the agglutinated form of כאשר):

(3.3)
הוא לא כשהיה [p2+k+vbp] פעם (כמו שהיה)
כשהייתי [p2+vbp] ילד אהבתי לשתות שוקו (כאשר הייתי)

Similarly, the prefix מש can be used as the sequence of מ and ש - in the sense of מאז ש- and sometimes as the sequence of מאשר:

(3.4)
הוא אוכל יותר משהוא [p2+k+pro] מדבר
משהגענו [p2+vbp] לבאר-שבע, רק נהננו

3.1.4 Suffix - suf

A *suf* tag is added to a word that has an accusative (for verbs), possessive (for nouns) suffix, or nominative (for prepositions, adverbs, and specific first-person present verbs.

(3.5)
עודני, אליו, להודיעם, ביתנו, חוששני

Inflections on אין *e.g.*, איננני, are not tagged as suffix.

3.1.5 Construct state - cons

A *cons* tag is added to the first word in a construct state (צירוף סמיכות).

(3.6)
ילדי [cons+nnms] הירח [p1+h+nnm (no cons tag)]

For numbers, the construct tag is also specified:

(3.7)
שלושת [cons+#] הילדים
שלושה [#] ילדים

3.2 Nouns

Common nouns in Hebrew can (in most cases) have a plural and a singular form, and can be modified with a possessive suffix, with a definite article ה or with של. Common nouns must be distinguished from proper nouns.

Positions, degrees and titles are tagged nouns:

(3.8)
רמ"ח, ד"ר, פרופ', יו"ר, רה"מ, ח"כת, אנג'

Nouns are modified by adjectives (the adjective agrees with the noun in number and gender), by prepositional phrases. They may be preceded by quantifiers (agree in number).
Nouns in dual form (e.g., ידיים) are marked as plural.

3.3 Adjectives

Adjectives are used to modify nouns. They agree in definiteness, gender and number with the noun. They can also appear in attributive sentences.

3.3.1 Adjectives vs. Nouns

Adjectives can sometimes be used with an implicit nominal head and fill the role of a noun in the sentence. In these cases, they are still tagged as adjectives.

- החכמים הכריזו. The token החכמים should be tagged as a definite noun, since the lexeme חכם is listed in the lexicon as a noun.
- התמימים נפלו קרבן להונאה. There is no noun entry for the lexeme תמים in the lexicon, so the token התמימים will be tagged as a definite adjective.
- הפרס נועד לפטור את המוכשרים מבעיות פרנסה. The token מוכשרים should be tagged as an adjective, due to the absence of noun entry for the lexeme מוכשר in the lexicon.

Even when the adjective is used in a construct state (*smiḥut*), and the overall phrase fills the slot of a noun phrase, the adjective is tagged as an adjective, *e.g.*, רעולי פנים חטפו אזרח בעיראק.

3.4 Verbs

Verbs express existence, action, or occurrence. Verbs in past, present, future tense and in the imperative agree in gender and number with the subject of the verb. Verbs in the future tense and feminine form - where a different masculine form exists for the same number - should be tagged with tags ending with *_f*. Verbs in the masculine form, or verbs for which only one form exists (e.g., verbs in the first person, singular, in past or future tense) should not be tagged as feminine, regardless of the context.

(3.9)
אני אנאם [vbf] מחר
לכו [vbi] מכאן
הן מתקדמות [vbr] עכשיו
הנערה אמרה: [vbp] אני אלך
הן תלכנה [vbf_f]
הן ילכו [vbf]

Verbs may (rarely) have an accusative suffix:

(3.10)
הוא הזיזם [vbp+suf] ממקום למקום

Verbs in infinitive form do not correspond in gender and number to the subject of the verb. They may (rarely) have an accusative suffix:

(3.11)
צריך לפעול [vb] מהר
יש להודיענו [vb+suf] בהקדם

Bare infinitives (צורת מקור) behave like nouns - they receive suffixes, can be modified by adjectives. But they cannot be modified by quantifiers and articles. In some idiomatic forms, a bare infinitive is used in a frozen construction. In this case, it is tagged as a noun.

(3.12)
שובו [vbb+suf] של הג׳דאי, *השוב שלו
שירי הלכת (frozen form - tagged as noun) [nn]

3.4.1 Auxiliaries

All inflections of the verb היה which precede verbs or modals are marked as regular verbs (and not as auxiliaries).

(3.13)
הייתי יכול להיות מיליונר

However, present-tense auxiliaries, which are actually copula, will be tagged as copula (see section 3.19):

(3.14)
החיים הם לא פיקניק

3.5 Beinoni

As noted by Gesenius [16, *p.* 355], *beinoni* [2] forms occupy a middle place between the noun and the verb. Morphologically, they are simple nouns, *i.e.*, they carry gender, number, and status inflections, definiteness, affixation, and no person and tense/mood relation. From the semantic point of view, according to traditional linguistics such as Gesenius [16], Hebrew participles are not representing a fixed state, but some source of action or activity, in contrast to nouns and adjectives (a claim which is not supported by nominalization). In [1] and [2] we introduce a distinct tag for *beinoni* forms specifically to avoid the systematic confusion that would otherwise occur between noun, adjective and verb tags. Our main motivation is that *beinoni* forms have specific syntactic features, which overlap only partially with each one of the major categories.

3.5.1 Present verbs vs. Participles

Present verbs has no construct state nor preposition prefix (ב כ ל מ). For the case of absolute state beinoni form which has no preposition prefix, the following tests can be used in order to to distinguish between present verb and participle.

Tense Affinity Participles have no tense affinity, in contrast to present verbs [6, *p.* 186].

[2] *Beinoni* and *participle* will be used in this document interchangebly

(3.15)
[verbal usage of beinoni form: present progressive]
החיילים מתאמנים עכשיו ⇐ החיילים מתאמנים אתמול*

(3.16)
[verbal usage of beinoni form: present simple]
החיילים מתאמנים בימים אלו ⇐ החיילים מתאמנים בימים ההם*

(3.17)
[nominal usage of beinoni form]
מתאמנים מגיעים ⇐ מתאמנים הגיעו

Explicit Subject Participles do not require an explicit subject, in contrast to present verbs. *e.g*, the token מטפסים in the phrase מטפסים הושקו can only be interpreted as a participle but not as a present verb.

3.5.2 Nouns vs. Participles

Several tests were suggested to distinguish between nouns and participles, as follows:[3]

1. Time-changing: a participle form which cannot be replaced by its past/future inflection is considered to be a noun or an adjective, otherwise, it is a preset verb.

(3.18)
צמח מטפס על הקירות ⇐ צמח טיפס על הקירות
מטפס הקירות פרח ⇐ טיפס הקירות פרח*

2. The genitive preposition של, can only precede a complement of a noun. Same for possessive pronoun suffix.

(3.19)
היא מנהיגה של קבוצה, היא מנהיגתם
היא מנהיגה את הקבוצה, היא מנהיגתם
*הוא לוכד של משרד החקלאות, הוא לוכדם
הוא לוכד את משרד החקלאות, הוא לוכדם

[3]The first test was suggested by Blao [6, *p.* 186], and the rest by Shlonsky [29, *pp.* 27–28].

3. Possessive construct state is possible only for nouns. The construct state of participles is always accusative.

(3.20)
שומרי המפעלים ⇐ השומרים של המפעלים
שומרי המפעלים ⇐ השומרים על המפעלים
לוכדי הנחשים ⇐ *הלוכדיםשל המפעלים
לוכדי הנחשים ⇐ הלוכדיםאת המפעלים

4. The prefix ה represents a definite article for nouns, and a relativizer for participles (*i.e.*, can be replaced by a ש relativizer).

(3.21)
השומר של המפעלים
השומר על המפעלים ⇐ ששומר על המפעלים

Note, that for this construction, quantification is not possible for a definite article, in contrast to relativizers.

(3.22)
*כל השומר של המפעלים יודע את תפקידו
כל השומר על המפעלים יודע את תפקידו

5. Transitive verbs with no complement should be interpreted as nouns, *e.g.* היא מנהיגה. On the other hand, a complement preceded by the accusative marker את is not possible for nouns: היא מנהיגה את הקבוצה.

6. An adjective modifier is only possible for nouns, *e.g.*, היא מנהיגה דגולה.

3.5.3 Adjectives vs. Participles

Doron [11] presents several tests to distinguish between adjective and (passive) participles.

1. The negation prefix בלתי modifies adjectives, *e.g.*, בלתי מוסמך vs. *בלתי משודר.

2. Words that can appear as complements of the verbs נראה, נותר are adjectives, *e.g.*, הוא נראה מסודר vs. *זה נראה מושר.

3. Participles of the form פעול are usually adjectives (except for those that are listed above), *e.g.*, כאוב, נעול.

4. When transforming the sentence to past or future, participle forms that function as adjectives do not change, and the auxiliary verbs היה, יהיה are added – האיש מסופר קצר ⇒ האיש היה מסופר קצר – under the same transformation, participle forms that function as verbs change to the proper tense – הסיפור מסופר בכל העיר ⇒ הסיפור יסופר בכל העיר.

5. Inversion is possible for verbs, but not for adjectives, *e.g.*, נגמר הפרויקט vs. *גמור הפרויקט.

6. The addition of על ידי is always possible for verbs, *e.g.*, מהלך מסורבל ⇒ *מהלך מסורבל על ידי המפלגה vs. העיר מותקפת ⇒ העיר מותקפת על ידי הצבא.[4]

7. Adjectives are gradable and can be modified by words such as יותר, הכי, *e.g.*, הוא מנהל מצליח ⇒ *הוא הכי מצליח vs. הוא המנהל הכי מצליח לנצח בכל משחק ⇒ הוא מצליח לנצח בכל משחק .

3.5.4 Participle adjectives vs. Participle nouns

In the case of a participle which is not listed as a noun or as an adjective, the POS may be determined by the role it fills in the sentence:

- קנינו שני מטפסים למרפסת – the token מטפסים is a participle noun.
- צמח מטפס – the token מטפס is a participle adjective.

However, we found such guidance to be complicated for the annotators, so the decision is to tag such forms as *participle*.

3.6 Adverbs

3.6.1 Adverbs in Modern Hebrew

According to Nir's discussion [24, chapter 18], adverbs in Modern Hebrew are the most heterogeneous set of all parts of speech.[5]. Adverbs are used to modify verbs of all forms, adjectives, quantifiers (especially those expressing inexact quantities), and full phrases, as follows:

[4] With some exceptions, such as החליפה תפורה על ידי חייט.

[5] For an overview of the evolution of adverb definition in MH, see [4, *pp.* 41–43].

- Verb: החולה מונשם מלאכותית.
- Adjective: הוא מנוסה פוליטית.
- Quantifiers: יחסית הרבה ילדים.
- Full phrases: חד משמעית, אין נזק בריאותי בטלפון סלולרי.

In contrast to other languages, such as English (*ly*) and French (*ment*), there is no one typical way or consistent method of adverb formation in Modern Hebrew. One can identify various types of adverb derivation:[6]:

- Closed list of 'pure' adverbs, mostly from the Bible [3, *p.* 594], *e.g.*, פתאום, די, חינם.
- *Conversion*, or *null derivation*, of adjectives and nouns to adverbs:
  - Singular masculine adjectives: שלושה פצועים קל, צריך לחשוב יצירתי, לכתוב נכון.
  - Plural feminine adjectives: הוא דיבר איתי גלויות, עניתי לו קצרות.
  - Nouns: אתמול, מחר,[7] סתם, כלום.
- Suffixation of ית *it* to singular nouns – פניתי אליו טלפונית – or ת to singular masculine adjectives[8] – הוא לחץ אותו אישית.
- Prefixation
  - ב with nouns[9]: במהירות, בהצלחה, בצחוק , בכתב.
  - ב with *beinoni*, mostly, in *pu'al* template which is, semantically, closed to adjective: במרומז, במיוחד.
  - ב with adjective, mostly in the Bible: בחזק יבוא, and in slang: ניצחנו בגדול.
  - כ ל מ with nouns: לבריאות, and adjectives: כמובן. As noted by Nir, this type of prefixation is used for derivation of adverbs of sentences, with pronoun suffix – לדעתי, לדאבוננו – and in parentheses – כנראה, לכאורה.
- Collocations

[6] See also [24, chapter 18], [3, *pp.* 593–601], [4, *pp.* 41–43].

[7] Definite nouns which denote period of time, *e.g.*, היום, הקיץ, can be considered as an expansion of such noun.

[8] Those adverbs can be looked at as adjectives in singular feminine form, with an omitted noun.

[9] As noted by Nir, this is the most common type of adverb formation.

- Lexical bound collocation, mostly from the canonic literature: קב, בעל פה ונקי,[10] as well as some loan translations: מעל ומעבר, דם קר, and phoneticisms: צ׳יק צ׳אק, חאפ לאפ.
- Free syntactic collocations, mostly based on the pattern: באופן/בצורה/בדרך/במידה + adjective, *e.g.*, במידה שווה.
- Binominals: יום יום, פרה פרה.

3.6.2 List of Adverbs

אדרבה, אולי, אומנם, אז, אזי, אחורה, אחורנית, אחרת, אט, איה, איך, איכה, איכשהו, אילך, איפוא, אישית, אכן, אכפת, אמנם, אמש, אנושות, אסור, אפוא, אפשר, ארוכות, ארצה, אשתקד, אתמול, באחרונה, באיבו, באין, באמצע, באמת, באריכות, בבירור, בגדול, בגילופין, בגלוי, בגנבה, בגניבה, בגסות, בגפו, בדד, בדומה, בדחילו, בדיוק, בדיעבד, בהגנב, בהדרגה, בהחבא, בהחלט, בהיות, בהכרח, בהצנע, בהקדם, בהקיץ, בהרבה, בהתאם, בהתאמה, בהתגנב, בואכה, בוודאות, בוודאי, בול, בזול, בזמנו, בחובו, בחוזקה, בחופזה, בחופשיות, בחוץ, בחזרה, בחטף, בחיוב, בחינם, בחשאי, בחשאיות, בטוחות, בטח, בטל, ביודעין, ביותר, ביחד, בייחוד, בינתיים, ביעף, בכדי, בכוונה, בכוח, בכך, בכלל, בכתב, בלאט, בלבד, בלעדית, בלעז, במאוחר, במדויק, במהירות, במהרה, במובהק, במופגן, במוצהר, במוקדם, במזיד, במטותא, במיוחד, במישרין, במכוון, בממוצע, במסתרים, במעורפל, במעט, במפגיע, במפורט, במפורש, במפתיע, במצודד, במקביל, במקצת, במקרה, במרוכז, במרומז, במשתמע, במתכוון, בנאמנות, בנוגע, בנוסף, בניחותא, בנפרד, בסדר, בסוף, בסיטונות, בסך, בעוד, בעין, בעיקר, בעליל, בעצלתיים, בעצם, בעצמו, בעקיפין, בערך, בפוטנציה, בפומבי, בפועל, בפירוש, בפנים, בפרהסיה, בפרוטרוט, בפרט, בצוותא, בצמצום, בקושי, בקיצור, בקירוב, בקפידה, בקצרה, בקרוב, בקשר, בראש, בראשונה, בראשית, ברבים, ברוגז, ברורות, בריבוע, ברצינות, בשמיים, בשנית, בתווך, בתוקף, בתכלית, גלויות, גם, גרידא, דאז, דאשתקד, דווקא, דום, דומם, די, דלקמן, דרומה, דרומית, האומנם, הבוקר, הביתה, הדדית, החודש, החוצה, היום, היטב, היישר, היתכן, הכי, הלאה, הלום, הלילה, הן, הנה, הסמסטר, העונה, הערב, הפעם, הצידה, הרבה, הרגע, הרחק, הרמטית, השבוע, השבת, השכם, השנה, השעה, ואילך, ובכן, וגו׳, ודאי, ודאית, וכד׳, וכדו׳, וכדומה, וכו׳, וכולי, ועוד, ותו, זמנית, חבל, חגיגית, חוץ, חזיתית, חזק, חינם, חיש, חלק, חלקית, חפיף, חרש, טוב, טלפונית, טרם, יומם, יותר, יחד, יחדיו, יחסית, ייתכן, ימה, ימינה, יפה, ישירות, ישר, כ, כאז, כאן, כביכול, כבעבר, כבר, כדאי, כדלהלן, כדלקמן, כה, כהוגן, כהלכה, כהנה, כיאות, כיום, כיצד, כך, ככה, כליל, כלכלית, כלל, כמדומה, כמה, כמובן, כמעט, כן, כנראה, כספית, כעת, כפליים, כראוי, כרגיל, כרגע, כשורה, לאחור, לאחרונה, לאט, לאלתר, לבד, לבסוף, לגמרי, להבא, להיום, להיכן, להיפך, להלן, לחוד, לחלוטין, לחלופין, לימים, לכאורה, לכאן, למדי, למהדרין, למחרת, למטה, למעלה, למעשה, למפרע, למצער, למקוטעין, למשעי, לנצח, לסירוגין, לעד, לעולם, לעולמים, לעיל, לעניין, לערך, לעתים, לפה, לפחות, לפנים, לפעמים, לפרקים, לפתע, לצדדים, לצמיתות, לצערי, לראשונה, לרבות, לרוב, לרעה, לשווא, לשעבר, לתמיד, מאוד, מאוחר, מאז, מאחור, מאידך, מאין, מאשתקד, מבחוץ, מדוע, מדי, מהיום, מהיכן, מהר, מוטב, מוסרית, מוקדם, מזמן, מזרחה, מזרחית, מחד, מחדש, מחר, מחרת, מחרתיים, מטה, מידית, מייד,

[10]Collocation modeling is supported by our representation – see section [1, section 6.3.1].

מינית, מכאן, מכבר, מלכתחילה, מלמעלה, מלפני, ממול, ממילא, ממש, מנין, מסביב, מספיק, מעולם, מעט, מעלה, מערבה, מעתה, מפה, מפורשות, מקומית, מקרוב, מר, מראש, משל, משם, מתי, נגבה, נורא, נחרצות, נכוחה, נכון, נכונה, נמרצות, נעים, סביב, סופית, סחורנית, סטטיסטית, ספישל, ספציפית, סתם, עדיין, עוד, עכשיו, עמומות, עמוקות, עקרונית, עתה, פה, פוליטית, פחות, פיקס, פנימה, פעם, פעמיים, פרא, פשוט, פתאום, צפונה, צפונית, קדורנית, קדימה, קודם, קלות, קמעה, קצת, קרי, קשה, קשות, ראשונה, ראשית, רב, רבות, ריאלית, רק, רשמית, שגרתית, שוב, שולל, שטוחות, שלשום, שם, שמא, שמאלה, שמה, שנית, שפי, תחילה, תיאורטית, תיכף, תכופות, תלוי, תמיד

3.6.3 Adverbs vs. Prepositional phrase

1. Adverbs, consisting of a preposition prefix and a noun, cannot be modified by an adjective:

(3.23)
הכוח נע בנחישות ⇐ הכח נע בנחישות אופיינית [prefix+noun]
הכוח פעל ברציפות ⇐ *הכח פעל ברציפות אופיינית [adverb]

2. Pronoun suffix is not common for nouns that compose adverb:

(3.24)
בנחישותו
*ברציפותו

The tagging criteria for tokens, consisting of a preposition prefix and a noun that modifies a verb, is based on the observation that adverbs cannot be suffixed by a pronoun nor be modified with an adjective. The token בנחישות, for instance, should be tagged as a noun with a preposition prefix, since it can be modified with an adjective הכוח נע בנחישות רבה or suffixed with a pronoun הכח נע בנחישותו האופינית. The token ברציפות, in contrast, will be tagged as an adverb, since no suffixation is possible: הכח נע ברציפותו האופינית* ⇒ הכח נע זה היום החמישי ברציפות.

3.6.4 Adverbs vs. Adjectives

An adjective that describes the situation of someone (or something) while performing an action is still an adjective, and not an adverb: הם התקדמו רועדים, תינוקת נמצאה משוטטת מיובשת. Adjectives, unlike adverbs, must agree with the noun or pronoun they modify in gender and number. If changing the gender or number of the subject of the verb calls for a change in the modifier – it is an adjective and not an adverb. In the following sentence, מהר is an adverb – הוא תייג מהר, while in the next sentence,

מתוח is an adjective – הוא חיכה מתוח לתוצאות, since it agrees in gender and number to the subject of the verb – הן חיכו מתוחות לתוצאות.

3.6.5 Adverbs vs. Verbs

In the following constructions, the verb fills the role of an adverb. It is still tagged as a regular verb or as a participle, since it is inflected by tense:[11]

(3.25)

מובן שיוסי הוא המנצח ⇐ יהיה מובן שיוסי הוא המנצח
יוסי, מסתבר, הוא המנצח ⇐ יוסי, הסתבר, הוא המנצח

Only lexicalized adverbs are tagged as adverb instead of prefixed verb, *e.g.*, כנראה, כמובן.

3.7 Conjunctions

Conjunctions in general establish a connection between two parts of a sentence. There exist three types of conjunctions:

Coordinating conjunction can join two entities (of the same syntactic category) that are equally important:

אבל, או, אולם, אז, אזי, אילו, אילולא, אד, אלא, אלמלא, אם, אף, אפילו, אפס, באם, באשר, ברם, גם, דהיינו, היינו, הלא, הן, הנה, הרי, ואכן, ובכן, כלומר, כן, לו, לולא, לכן, למשל, לפיכך, רק, בין

Subordinating conjunction comes at the beginning of a subordinate clause and establishes a relationship between the dependent clause and the rest of the sentence. The dependent clause meaning depends on the rest of the sentence and does not stand on its own:

פן, כדי, בלי, במידה-ש, בעוד-ש, ברי-ש הגם-ש, הואיל-ו, הודות-ל, היות-ש, היות-ו, כאילו, כאשר, כדי, כי, כיוון, ככל-ש, כפי-ש, כש, כשם-ש, לאחר, לבל, למרות-ש, מאחר-ש, מאחר-ו, מבלי, מכדי, מכיוון-ש, מכפי, מפני-ש, משום-ש, שכן, שמא, בטרם, בגלל-ש, בלא, בלא-ש, בשביל-ש, כמו-ש, כמות-ש

[11] Note that some of these forms have an adverb translation in English.

Relativizing conjunction connects a clause to a noun phrase:

ש, אשר

3.8 Prepositions

3.8.1 Prepositions in Modern Hebrew

Ben-Asher [5] discusses the definition and the syntactic role of prepositions in Modern Hebrew. He argues against definitions which are based only on semantic criteria, such as Goshen *et al.* [17, *p.* 4], Nahir [22, *p.* 7], and Livny & Kochvah [18, *p.* 97], as well as against Yo'eli [33], Blao [6], and Zadka [35], who involve syntactic considerations such as word ordering and exclude morphological criteria.

Ben-Asher's preposition definition is based, mainly, on morphological criteria with syntactic and semantic considerations:

1. In contrast to nouns, only pronominal pronouns can be attached to prepositions, but not independent pronouns (subjective mode), *e.g.*, אצל אני* אצלי, בגלל אתה* בגללך, *vs.*, עוד אני / עודני.

2. There is no plural form for prepositions.

3. The attachment of a pronominal pronoun to a given preposition is based on either singular or plural baseform, but not on both of them: אלי אליך אליו ⇒ *אלי אלך אלו, נגדי נגדך נגדו ⇒ *נגדי נגדך נגדיו. As noted by Ben-Asher, this rule does not apply to the third-person suffixation of the preposition בין, which has both singular-based בינם and plural-based ביניהם suffixations.

4. Prepositions can precede only a noun or a pronoun. Some prepositions – אחרי, לפני, מן, עד – may come before adverbs, *e.g.*, עד שלשום, לפני אתמול, אחרי מחר, מן היום.

5. Semantic considerations, such as the relation between words, may be taken into account in addition to the above criteria.

As for the syntactic role of the prepositions, Ben-Asher follows Brockelmann [7] in considering prepositions as nouns in a descriptive role which starts an adverbial phrase, and where the preposition is the head of the construct state and the noun is the modifier. In some cases, prepositions start indirect object or modifier phrases.

Most of the prepositions can start subordinate clauses[by adding the the ש morpheme, *e.g.*, לפני ש. Some of the prepositions may fill the role of conjunction, with no additional morpheme – למען, בעוד, עקב, בעבור, טרם, מאז – or by adding ש morpheme – בגלל ש, על אף ש, למרות ש.

3.8.2 List of Prepositions

אגב, אודות, אחר, אחרי, אל, אצל, את, ב, באמצעות, בבחינת, בגדר, בגין, בגלל, בדבר, בואכה, בזכות, בטרם, ביד, בידי, בין, בכלל, בלא, בלוויית, בלי, בלית, בלעדי, בלת, במו, במחיצת, במעין, במקום, במשך, בנוסף-ל, בניגוד-ל, בעבור, בעד, בעוד, בעזרת, בעטי, בעיצום, בעקבות, בפני, בצד, בקרב, בקשר-ל, בשביל, בשל, בתוך, בתוככי, בתור, דרך, זולת, חוץ-מ, חרף, כאל, כגון, כדי, כולל, כלפי, כמו, כמות, כמין, כמעין, כנגד, כעבור, כעין, ל, לאור, לאחורי, לאחר, לבד-מ, לבין, לגבי, לזכות, ליד, לידי, לכבוד, לכדי, לכעין, ללא, למן, למעט, למעין, למען, למרות, לנגד, לנוכח, לעבר, לעומת, לפי, לפני, לצד, לקראת, לרגל, לשל, לשם, לתוך, מאחורי, מאחרי, מאל, מאת, מבין, מבלי, מבעד, מול, מחד, מחמת, מטעם, מידי, מלבד, מן, מעבר, מעין, מעל, מפני, מצד, מקרב, משל, מתוך, מתחת, נגד, נוכח, סביב, עבור, עד, על, עם, עקב, פן, פרט-ל, קבל, קודם, קרוב-ל, של, תוך, תחת, תמורת

3.8.3 Suffixed preposition vs. Prefixed pronoun

Tokens such as בם could be tagged either as a preposition **baseform** ב followed by a pronominal suffix הם, or as a preposition **prefix** ב followed by a pronominal suffix הם, with no baseform stem. We chose the first option, since it is consistent with the case of prepositions which cannot be agglutinated, *e.g.*, עליו – there is no prefix על.

3.8.4 Preposition vs. Conjunction

1. Prepositions which are followed by a noun phrase, and are not used to be followed by ש:

   - ניצחנו על אף השיפוט הביתי – the collocation על אף is a preposition.
   - ניצחנו אף על פי שהשיפוט היה ביתי – the collocation אף על פי is a conjunction.

2. Conjunctions of coordination – ו, או – can bind any type of syntactic categories and phrases. Prepositions are always followed by a noun phrase:

   - הוא חלה עקב הזנחת בריאותו – the word עקב is a preposition.
   - הוא חלה, כיוון שהזניח את בריאותו – the word כיוון is a conjunction.

3.8.5 Preposition vs. Prefixed Noun

Some preposition words could be interpreted as a noun with an agglutinated preposition, *e.g.*, מצד, מפני, מחמת, מטעם, במלאות, במשך, בתום, בעקבות, בזכות. The test to determine whether the token is used as a preposition (single word) or as a noun prefixed by a preposition, is the following:

1. If the word של can be inserted, the word is used as a noun:
   - מטעם ראש הממשלה ⇒ *מטעם של ראש הממשלה – the token מטעם is a preposition.
2. If a quantifier can be inserted before the noun, the word is used as a noun:
   - ברחבי העולם ⇒ בכל רחבי העולם – the token ברחבי is a prefixed noun.
   - בעקבות עליית מחירי הנפט ⇒ *בכל עקבות עליית מחירי הנפט – the token בעקבות is a preposition.

3.9 Quantifiers and Determiners - DT,QR

Quantifiers and determiners are used to quantify or narrow the reference/identity of the denoting noun phrase (NP) in various manners. These words are generally located before the head of the NP.

3.9.1 Quantifiers vs. Determiners

Quantifiers (*e.g.*, מספיק, כמה, קצת, די, פחות, מעט, מחצית, כל, יותר) and determiners (*e.g.*, אף , מין, אותו, איזה, מרבית, רוב, עיקר, מיטב, כל, מבחר, שאר, יתר) are sometimes distinguished (see the structural/morphological criteria proposed in [10] and [32]):

1. Quantifiers can be used in partitive constructions using מ- and מן prepositions:

   (3.26)
   שלושה מתלמידיו
   *רוב מתלמידיו

2. Quantifiers can determine the agreement features of the NP (determiners can't):

(3.27)
בקבוק סודה אחד עולה שקל
*מרבית הבחורות אוהבות את בן זוגה

3. Quantifiers can be modified or quantified themselves, but determiners can never be quantified nor quantified:

(3.28)
פי שניים מבחר הקצינים
הרבה מאד ילדים
*המיטב הרב של הילדים

4. A phrase of the form Qr+יש לנו is valid while Dt+יש לנו is not:

(3.29)
יש לנו מחצית/שלושה/יותר/רבע/מעט
*יש לנו מרבית/עיקר/שאר/אותם/מבחר/כל/אף

5. There will be at most one quantifier in a NP but it is possible to have more than one Determiner, which will always precede the quantifier:

(3.30)
מרבית אותם ילדים
כל עשרת הדברות

A determiner cannot be quantified by itself but is recursive in structure and always precedes the quantifier. A determiner will always be in construct state while for a quantifier it is not a must.

6. Determiners are indefinite, in contrast to quantifiers:

(3.31)
כל ילד בכיתה צריך להביא ספר [DT]
כל הילדים בכיתה הכינו שעורי בית [QR]

Possible Confusions Determiners and quantifiers positioned after the head behave more like adjectives: they agree in number, gender and definiteness with the head. Determiners with the same meaning can appear before or after the head, as in the following examples. In all cases, these words are tagged as DT:

(3.32)
כל הילדים - הילדים כולם

In some cases, only one of the two options is available: *הילדים הרבים - הרבה הילדים since when appearing before the head, הרבה is lexically marked as indefinite, and therefore cannot be used with a definite head.

(3.33)
הבנות כולן
הבנים רובם
ילדים רבים

3.9.2 Practical Decision - One tag (DT) for Quantifiers and Determiners

We tag determiners and quantifiers consistently as DT since the distinction can be difficult for human taggers, and can be performed as a post-processing specialized task.

No construct state (cons) should be denoted for determiners.

3.9.3 List of Quantifiers/Determiners

We use one tag QR for three types of quantifiers, as follows:

- Amounts

  עוד, הרבה, המון, יותר, כמה, כפול, פחות, די, מספיק

  - Amount quatifiers are indefinite, *e.g.*, *היותר אנשים.
  - The noun which they quantify is indefinite, *e.g.*, *יותר האנשים.
  - Can be set in partitive structures, *e.g.*, יותר מהאנשים לא יבואו.
  - They are elliptic, *e.g.*, היו לי שלושה שקלים, עכשיו יש לי יותר.
  - Gender/number/person inflection is not possible.

  - Absolute state is possible for partitive structures.
  - Nominative suffix is not possible, *e.g.*, יותרם*.

  In addition, the following quantifiers is classified as *amounts*: מקצת, קצת, מרבית, מעט

- Partitives

  רוב, יתר, שאר, כל, כלל, מירב, מיטב

  - Definite article can attach partitives, *e.g.*, הכל, הרוב.
  - The noun which they quantify is definite, *emph*, רוב הילדים, רוב ילדים*.
  - Cannot be set in partitive structures, *e.g.*, רוב מהאנשים לא יבואו*.
  - A sequence of partitives is possible, *e.g.*, כל שאר הילדים.
  - Gender/number/person inflection is not possible.
  - Partitives are always in construct state.
  - Nominative suffix is possible for some partitives, *e.g.*, כולן, רובו, כללם.

  In addition, the following quantifiers is classified as *partitives*: מספר, מבחר

- Determiners

  אותו, איזה, עיקר, עצם, כל, מין, אף, שום, מידי

  - Determiners are indefinite.
  - Can be classified by two paranmeters: gender/number/person inflection (+/-GNP), possible definiteness of the noun it quantifies (+/-D):
    * +GNP,+D: אותו
    * +GNP,-D: איזה
    * -GNP,+D: עיקר, עצם
    * -GNP,-D: כל, מין, אף, שום, מדי

3.10 Numbers -

We tag as Numbers (#) every occurrence of a numeral - in words, in digits or in alphabetical ('gematria'),א' באדר יא including those which function as a quantifier.

(3.34)
שלוש בנות (before noun)
לולאות חמישים (after noun)
שלושה באו (alone)
שלושה מהילדים (partitive)
השלישי יקבל מדליית ארד

3.11 Existentials

EX - יש אין
יש is used as the present tense of the verb היה in cases of possessive, existential, locative clauses. Though יש does not fully behave as a verb, we tag it as an existential auxiliary EX when it is not used as a modal.

(3.35)
יש/היה/יהיה לי המון כסף בבנק (possession)
יש מילים קטנטנות, חמודות, כמו: אם (existential - universal)
יש/היו/יהיו מלפפונים במקרר (existential - locative)

3.11.1 Existential (possessive) vs. Copula

In the following examples, הם is used as a copula, whereas היו is used as an exsitential[12]:

(3.36)
הבולים האלה הם שלי (copula)
היו לי בולים (existential)

3.12 Negations - NEG

The following words belong to this category: לא, לאו, אל, בל, אין

[12]see also 3.19.2

(3.37)
הוא לא נמצא
אל תלכי לשם
לאו דווקא
בל תוסיף

3.12.1 Negation vs. Copula

Pronominal suffixed אין (*e.g.*, אינה ,אינו) should always be tagged as copula (see 3.19).

3.12.2 Negation vs. Existential

Negative אין can be replaced with לא:

(3.38)
אין הנחתום מעיד על עיסתו ⇐ לא הנחתום מעיד על עיסתו [NEG]
אין חיים על הירח ⇐ *לא חיים על הירח [EX]

Existential אין can be replaced with יש

(3.39)
אין חיים על הירח ⇐ יש חיים על הירח [EX]
אין הנחתום מעיד על עיסתו ⇐ *יש הנחתום מעיד על עיסתו [NEG]

3.13 Modals - MD

Modals express the speaker's opinion on a fact, deontic (must) and epistemic (may, can, might) modality. While in English modals have clear-cut syntactic properties (modal auxiliaries are a well-defined closed class), there is no similar category in Hebrew.

In [23] and [1] we reviewed three major approaches to categorizing modals in Hebrew:
Semantic – represented mostly in Kopelovich's work, modality is categorized by three dimensions of semantic attributes. Since her claim is that there is no syntactic category of modality at all, this approach 'over-generates' modals and includes words that from any other syntactic or morphologic view fall into other parts of speech.
Syntactic-semantic – Zadka classifies seven sets of verbs and pro-verbs following syntactic and semantic criteria. His claim is that modality actually is marked by syntactic characteristics, which can be identified by structural criteria. However, his

evaluation mixes semantics with syntactic attributes.
Morphological-syntactic – Rosen's definition of *Impersonals* is strictly syntactic-morphological and does not try to characterize words with modality. Consequently, words that are usually considered modals are not included in his definition, such as אסור *'asur* (forbidden), מותר *mutar* (allowed), יכול *yakol* (can).

The variety of criteria proposed by linguists reflects the disagreements we identified in lexicographic work about modal-like words in Hebrew. We introduce a modal tag in our Hebrew tagset. Although there is no distinct syntactic category for modals in Hebrew, we propose the following criteria:

1. They have an infinitive complement or a clausal complement introduced by the binder ש *š*.
2. They are NOT adjectives.
3. They have irregular inflections in the past tense, *i.e.*, רציתי לדעת *raciti lada'at* (I wanted to know) is not a modal usage.

3.13.1 List of Modals

- Uninflected modal lexemes

  איכפת, אין, אפשר, אסור, די, יש, כדאי, חבל, חובה, מוטב, מותר, אל, בל

- Gender-numer inflected modal lexemes

  אמור, מוכן, מסוגל, ניתן, עלול, עשוי, רשאי, זכאי, חייב, חשוב, צריך, ראוי, זקוק, מוכרח, רצוי, ניתן

- Gender-numer-tense inflected modal lexemes

  יכול

3.13.2 Modal vs. Existentials

The words יש and אין can be used either as existential verbs or as modals:

(3.40)
אין לאשר את המסמך ללא נוכחות המנכ"ל [MD]
אין לאשר מושג במתמטיקה [EX]
יש לחכות עד גמר הנחיתה [MD]
יש לחכות קרס מיוחדת [EX]

The tests to distinguish existential and modal readings of יש and אין are:

1. EX can be replaced by לא היה/לא יהיה
2. MD can be replaced by צריך

3.13.3 Modal vs. Adjective

Some adjectives accept infinitival complements in impersonal constructions:

(3.41)
נעים לטייל בחורף
קל להרוס

The tests to distinguish this usage from a modal usage are:

1. Adjectives are gradable and can be modified by מאוד or יותר
2. Adjectives can become describers of the nominalized verb:

(3.42)
נעים לטייל בחורף ⇐ הטיול נעים מאוד
קל להרוס ⇐ ההריסה קלה מאוד
יוסי מסוגל להרים את המכשיר ⇐ *הרמת המכשיר ע"י יוסי מסוגלת מאוד

3.14 Proper nouns - PNM

Proper nouns are names of persons, locations (cities, states) and corporations. Proper nouns generally refer to singular entities, and hence they are not modified with a definite article or have a plural form: צחי הנגבי, איראן, ירושלים, יובנטוס, אינטל.

Names of organizations are often acronyms and are tagged as proper nouns: צה"ל, ש"ס.

Proper nouns used with definite articles are marked in the lexicon. They are tagged [p1 PNM]. They correspond in general to the following classes:

- Names of rivers: הירדן, הירקון, החידקל
- Name of regions: הגולן, הגליל, השפלה, הנגב, הותיקן
- Name of mountains: הגולן, החרמון

- Acronyms that name organization: השב"כ האו"מ הרש"פ
- Names of newspapers: הארץ, המודיע, היום
- Names of political parties: הליכוד, העבודה, המערך, החיזבאללה

3.14.1 Proper noun vs. Noun

1. Noun phrases consisting of words that are not proper names should be tagged as ordinary noun phrases even if the full phrase is the name of an organization.

(3.43)
שרות [cons+nnm] הביטחון [p1+h+nnm] הכללי [p1+jjm]
פיקוד [cons+nnm] צפון [nnm]

2. Names of models (of cars, airplanes, etc.) should be tagged as proper nouns even if, in their given context, they are modified with a definite article or have a plural form. Semantically, in those cases the modal name refers to a particular object of that type.
Here קלאצ'ניקוב is a proper noun:

(3.44)
הוברחו שלושים רובים מסוג קלאצ'ניקוב

In the following sentence קלאצ'ניקוב is in plural form, still tagged as a proper noun:

(3.45)
הוברחו שלושים קלאצ'ניקובים

In the sentence, וולבו is a proper noun that refers to the name of corporation:

(3.46)
עליית במכירות וולבו בישראל

In the following, וולבו is used as an instance of a class, it is still tagged as a proper noun:

(3.47)
קניתי)את ה(וולבו לפני חודש

3.15 Interrogatives - INT

This category includes question words:

- Pro-noun

  מה, מי

- Pro-adverb

  איה, איך, אימתי, איפה, אנה, היכן, כיצד, לאן, למה, מדוע, מתי

- Pro-determiner

  איזה, איזו, איזהו, איזוהי, אילו, כמה

- Yes/No

  האמנם, האם, הגם, הייתכן

3.16 Pronouns - PRO

Pronouns appear in the following forms:

- Personal pronouns

  אני, אנוכי, אנחנו, אנו, אתה, את, אתם, אתן, הוא, היא, הם, הן

- Impersonal pronouns

  כמה, כלום, כולם, מישהו, כלשהו, משהו

- Demonstratives pronouns

  זה, זו, זהו, זוהי, זאת, ההוא, ההיא, ההם, ההן, אלה, אלו, הלה, האלה, האלו, הללו, הזה, הזו, הזאת, הלז, הלזו, הלזה, כך

- Reflexive pronouns

עצמי, עצמנו, עצמך, עצמכם, עצמכן, עצמו, עצמה, עצמם, עצמן

Personal pronouns are marked in person, number, gender (except for first person) as well as in case (and will be agglutinated for any case other than subjective).

Personal pronouns can replace whole noun phrases:

(3.48)
הילד הקטן סיפר לי סיפור ⇐ הוא סיפר לי סיפור

Pronouns cannot be modified:

(3.49)
*הוא היפה בא לבקר אותי

Demonstrative which appear in a noun phrase agree in number, gender, and definiteness with the noun phrase head:

(3.50)
ילד זה
הילד הזה
הילדה הזו
ילדים אלה
הילדים האלה

3.17 Prefix words - PREF

Prefixes are words that are have no independent status, and always appear immediately before some other word (sometimes separated with a '-'):

אי, אין, אנגלו, אנטי, אקס, בין, בלתי, בן בר, בתר, דו, דמו, חד, חוץ, טרום, כלל, לוקאל, לטינו, מגה, מולטי, מיני, נאו, נון, סמי, על, פאן, פוסט, פסאודו, פרא, פרה, פרו, קדם, רב, תוך, תלת, תת

Prefixes have no inclinations or plural/gender form:

(3.51)
חלקיק תת-אטומי ⇐ חלקיקים תת-אטומיים
מוצר אנטי בקטריאלי ⇐ משחה אנטי בקטריאלית

When used without a continuing noun or adjective, the prefix is tagged as a noun:

(3.52)
קניתי בסופר

3.18 Foreign words - FW

This tag is for words in a foreign language, in Hebrew or English letters, which are not proper nouns.

(3.53)
הנשיא אמר: ריד מיי ליפס, לא אטיל מיסים חדשים
הפסדנו את הגביע, מה לעשות, סה לה וי

3.19 Copulas - CP

Copulas inflect for number, gender, person and polarity, with *positive* and *negative* as its values. Lexical items in this category include:

- הנני, הנך, הנך, הינו, הינה, הוא, היא
- הננו, הנכם, הנכן, הינם, הינן, הם, הן
- איננני, איננך, איננך, איננו, איננו, איננה
- איננה, איננו, איננכם, איננכן, איננם, איננן
- היה and its inflections
- יהיה and its inflections

Semantically, copulas can express

- Equative relations

(3.54)
אריק הוא רופא
אריק הוא הרופא

- Attributive relations:

(3.55)
אריק הוא אמיץ

- Locative relations:

(3.56)
אריק הוא בבית
המסיבה היתה אתמול

3.19.1 Copula vs. Pronoun

היא/הוא can be tagged as pronoun or copula:

1. Pronoun can be replaced by אני, אתה etc.

(3.57)
הוא רוצה לשתות ⇐ אתה רוצה לשתות

2. Copula: (a) can be replaced in the past by היה (b) can be negated as אינו

3. Pronoun can be replaced by אני, אתה etc.

(3.58)
אריק הוא אמיץ ⇐ אריק היה אמיץ
אריק הוא אמיץ ⇐ אריק אינו אמיץ

Special case:

(3.59)
התמונה היא היא המציאות שבה גדלנו

This is an intensive variation of the copula היא - we would tag it as a 2-word copula.

3.19.2 Copula vs. Existential

היה and יהיה can be tagged as copula or existential.

1. Copula (a) can be omitted in the present (b) can be replaced by הוא in the present.

(3.60)
אריק היה אמיץ ⇐ אריק אמיץ
אריק היה אמיץ ⇐ אריק הוא אמיץ

When היה is used to mark the tense or aspect of a verb that appears in participle form - tag it as a copula as well:

(3.61)
אריק היה הולך כל בוקר לביתו

2. Existential can be replaced by יש in the present

(3.62)
היו 3 שולחנות בחדר ⇐ יש 3 שולחנות בחדר

Bibliography

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

[2] Meni Adler, Yael Netzer, Yoav Goldberg, David Gabay, and Michael ELhadad. Tagging a hebrew corpus: The case of participles. In *Proceeding of the 6th edition of the Language Resources and Evaluation*, Marrakech , Morocco, 2008.

[3] Isaac Avinery. *YAD HALLASHON - Lexicon of Linguistic Problems in the Hebrew Language*. Yizra'el, Tel Aviv, Israel, 1964. (in Hebrew).

[4] Mordechai Ben-Asher. *The Consolidation of the Normartive Grammar*. Hakibbutz Hameuchad, Haifa, Israel, 1969. (in Hebrew).

[5] Mordechai Ben-Asher. On the prepositions in Mosern Hebrew. *Lešonenu*, XXXVIII:285–294, 1974. (in Hebrew).

[6] Yehoshua Blao. *Syntax Fundamentals*. Hebrew Institute for Written Education, Jerusalem, 1966. in Hebrew.

[7] Carl Brockelmann. *Grundriss der vergleichenden Grammatik der semmitischen Sprachen*. Georg Olms Verlagsbuchhandlung, Hildesheim, 1966. (in German).

[8] David Carmel and Yoelle S. Maarek. Morphological disambiguation for Hebrew search systems. In *Proceeding of NGITS-99*, pages 312–326, 1999.

[9] Yaacov Choueka. *Rav-Milim - A Comprehensive Dictionary of Modern Hebrew, literally: Multi-Words*. C.E.T, Miskal and Steimatzky, Tel-Aviv, Israel, 1997.

[10] Edit Doron. The np structure. In Uzi Ornan, Edit Doron, and A. Ariely, editors, *Hebrew Computational Linguistics*. Ministry of Science, 1991. (in Hebrew).

[11] Edit Doron. The passive participle. *Hebrew Linguistics*, 47:39–62, 2000. (in Hebrew).

[12] Oswald Ducrot and Tzvetan Todorov. *Encyclopedic dictionary of the science of language*. John Hopkins University, Baltimore, MD, 1979.

[13] Knowledge Center for Processing Hebrew. Hebrew morphological analyzer. http://mila.cs.technion.ac.il.

[14] W. N. Francis. A tagged corpus - problems and prospects. In S. Greenbaum, G. Leech, and J. Svartvic, editors, *Studies in English Linguistics for Randolph Quirk*, pages 192–209. Longman, London and New York, 1979.

[15] Roger Garside, Geoffrey Leech, and Geoffrey Sampson. *The computational analysis of English. A corpus-based approach*. Longman, London, 1987.

[16] Friedrich H. W. Gesenius. *Hebrew Grammar*. The Clarendon Press, Oxford, 1976. Edited and enlarged by E. Kautzsch, English edition by A. E. Cowley.

[17] Moshe Goshen-Gotshtein, Ze'ev Livne, and Shlomo Shpan. *The Practical Hebrew Grammar*. Schoken, Jerusalem, 1966. in Hebrew.

[18] Yitschak Livny and Moshe Kochva. *Hebrew Grammar*. 'ever, Jerusalem, 1965. in Hebrew.

[19] Ralph B. Long. *The Sentence and its Parts*. University of Chicago Press, Chicago and London, 1961.

[20] Christopher D. Manning and Hinrich Schutze. *Foundation of Statistical Language Processing*. MIT Press, 1999.

[21] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marchinkiewicz. Building a large annotated corpus of English: The penn treebank. *Computational Linguistics*, 19:313–330, 1993.

[22] Simcha Nahir. *The Principles of the Senetnce Theory*. The Hebrew Realistic School, Haifa, 1963. in Hebrew.

[23] Yael Netzer, Meni Adler, David Gabay, and Michael ELhadad. Can you tag the modal? you should! In *ACL07 Workshop on Computational Approaches to Semitic Languages*, Prague, Czech, 2007.

[24] Raphael Nir. *Word-Formation in Modern Hebrew*. The Open University of Israel, Tel-Aviv, Israel, 1993.

[25] Uzi Ornan. The parts of speech. *Lěšonénu*, XXV:35–56, 1960. (in Hebrew).

[26] Haim B. Rosen. *Contemporary Hebrew*. Mouton, The Hague, Paris, 1977.

[27] Ora (Rodrigue) Schwarzald. *Studies in Hebrew Morphology*. Volumes 1–4. The Open University of Israel, Tel-Aviv, Israel, 2002. (in Hebrew).

[28] Erel Segal. Hebrew morphological analyzer for Hebrew undotted texts. Master's thesis, Technion, Haifa, Israel, 2000. (in Hebrew).

[29] Ur Shlonsky. *Clause Structure and Word Order in Hebrew and Arabic*. Oxford Universirt Press, New York Oxford, 1997.

[30] Khalil Sima'an, Alon Itai, Alon Altman Yoad Winter, and Noa Nativ. Building a tree-bank of modern Hebrew text. *Journal Traitement Automatique des Langues (t.a.l.)*, 2001. Special Issue on NLP and Corpus Linguistics.

[31] James Sledd. *A Short Introduction to English Grammar*. University of Texas, Scott, Foresman and Company, 1959.

[32] D. Yizhar. Computational grammar for noun phrases in Hebrew. Master's thesis, Hebrew University, 1993. (in Hebrew).

[33] M. Yoeli. *Hebrew Syntax*. Yesodot, Tel-Aviv, 1963. in Hebrew.

[34] Shlomo Yona. A finite-state based morphological analyzer for Hebrew. Master's thesis, Haifa University, 2004.

[35] Yitzhak Zadka. *The Practical Hebrew Grammar*. Qiryat Seper, Jerusalem, 1995. (in Hebrew).