Tagging a Hebrew Corpus: The Case of Participles

Meni Adler, Yael Netzer, Yoav Goldberg, David Gabay and Michael Elhadad

Ben Gurion University of the Negev
Department of Computer Science
POB 653 Be’er Sheva, 84105, Israel
{adlerm|yaeln|yoavg|gabayd|elhadad}@cs.bgu.ac.il

Abstract

We report on an effort to build a corpus of Modern Hebrew tagged with parts of speech and morphology. We designed a tagset specific to Hebrew while focusing on four aspects: the tagset should be consistent with common linguistic knowledge; there should be maximal agreement among taggers; and the tagset should be effective for applications relying on the tags as input features. In this paper, we illustrate these issues by explaining our decision to introduce a tag for beinoni forms in Hebrew. We explain how this tag is defined, and how it helped us improve manual tagging accuracy to a high-level, while improving automatic tagging and helping in the task of syntactic chunking.

1. Introduction

This paper discusses decisions taken during our work in establishing a tagset for Hebrew. The method we adopted for this purpose aims to find a tagset that maximizes agreement among taggers but maintains maximal consistency with morphological characteristics of the words, and consequently with traditional perceptions of syntactic, semantic and lexical resources.

One of the main issues relevant when tagging Semitic languages is that the orthographic form of words allows for agglutination of prefixes and suffixes into a single token. Taggers for Hebrew as described in Adler and Elhadad (2006) assume a word-based model, the tagset we will design must consequently be word-oriented – that is, we expect the tags to describe full words as opposed to separate morphemes. In this paper, we focus on the case of beinoni forms in Hebrew. The issue is how beinoni should be tagged. We consider three main approaches: treat beinoni (participle) forms as either verbs, nouns and adjectives according to the context; treat beinoni forms as verbs; or – the approach we adopt – add a participle tag to the tagset. Existing lexical resources do not include such a participle category. We show that these resources exhibit high disagreement on the POS they predict for beinoni forms, which causes inconsistencies in tagging. In contrast, using the guidelines we designed, taggers achieved a very high level of agreement. We also discuss how the presence of the participle tag affects tasks that depend on the tagged corpus, such as syntactic chunking.

2. Corpus and Tagging Process

In recent years, two large-scale computational resources have been developed for Hebrew as part of the Hebrew Knowledge Center initiative: a corpus compiled and manually tagged at Ben Gurion University, and the Hebrew Treebank generated at the Technion (Sima’an et al., 2001). Tagging in the treebank project is syntax-oriented, while in the tagged corpus we describe here, the approach is lexicon-oriented: a lexicon of Hebrew words proposes for each word a list of possible tags, and the tagged corpus indicates the correct tag in context.

One of the main objectives we assigned to ourselves while developing this corpus, was to design a specific tagset appropriate for Hebrew. We did not assume a priori that an existing tagset (adopted from English or from traditional dictionaries) would be appropriate to fulfill the requirements on a high-quality computational corpus. Our first objective is to maximize agreement among human taggers, in order to ensure consistency of the tagged corpus.

However, agreement among taggers cannot be our only criterion for tagset quality, otherwise the trivial tagset of one tag (WORD) would be optimal – but non-informative. Most meaning-carrying words belong to one of the main three categories (verbs, nouns and adjectives). Taggers achieved above 70% agreement (between 4 people) on the very first training round while focusing on these three base categories.

In a minimalistic approach, we would adopt the following heuristic: define an OTHER category for all the words where no clear-cut agreement on a category can be reached (if a word is not a clear and well-behaved verb, noun or adjective - tag it as OTHER). We found that such a method did not increase agreement in any way. In addition, this approach would have also caused bad learning of a stochastic model of context. It is critical to model words such as prepositions or conjunctions to correctly disambiguate verbs or nouns.

One of the main confusing factors we found among taggers was related to the status of what we call beinoni. The main reason is that the beinoni form (literally the “middle form” of verbs) is a form that shares morphological and syntactic properties of nouns, verbs and adjectives. We explain below our decision to introduce a distinct participle tag in the tagset, and present the guidelines we have designed to define it.

Our corpus is comprised of short news stories. It includes about 40M tokens, in articles of length between 200 and 1,000 tokens. Of the full corpus, a sample of articles comprising altogether 200K tokens was assembled at random,
and manually tagged for part of speech by four taggers (for details see Adler (2007, chapter 4)). An initial set of guidelines was first composed, relying on the categories found in several dictionaries and on the Penn treebank POS guidelines (Santorini, 1995). As many words from the corpus were either missing or tagged in a non uniform manner in the lexicons, we recommended looking up missing words in traditional dictionaries. However, disagreement was found in these dictionaries, among traditional dictionaries, both for open and closed set categories. Given the lack of a reliable lexicon, the taggers were not given a list of options to choose from, but were free to tag whatever tag they found suitable.

Initially, each text was tagged by four people, and, iteratively, the guidelines were revised according to questions or disagreements that were raised. As the guidelines became more stable, the disagreement rate decreased, each text was tagged by three people only and eventually by two taggers and a referee that reviewed disagreements between the two. The disagreement rate between any two taggers was initially as high as 20%, and dropped to 3% after a few rounds of tagging and guidelines revision. Initially, each text was tagged by four people, and, iteratively, the guidelines were revised according to questions or disagreements that were raised. As the guidelines became more stable, the disagreement rate decreased, each text was tagged by three people only and eventually by two taggers and a referee that reviewed disagreements between the two. The disagreement rate between any two taggers was initially as high as 20%, and dropped to 3% after a few rounds of tagging and guidelines revision.

Major sources of disagreements include, preposition phrases, adverbial phrases, modals (Netzer et al., 2007) and beinoni. We focus in this paper on beinoni forms. Beside the disagreement among taggers, we also found significant disagreement among Modern Hebrew dictionaries. Table 1 lists the various selected POS tags for words we identify as beinoni form, as determined by: (1) Rav Milim (Choueka et al., 1997), (2) Sapir (2002), (3) Even-Shoshan (2003), (4) Knaani (1960), (5) HMA (Carmel and Maarek, 1999), (6) Segal (2000), (7) Yona (2004) and (8) the Hebrew Treebank (Sima’an et al., 2001). As can be seen, there is almost systematic confusion between Verb, Noun and Adjective tags for these words. We propose guidelines which remove this confusion, and allowed us to reach very high agreement among taggers. We also discuss how the new ‘participle’ tag we introduce is used by a syntactic chunker.

3. Previous Work
The question of which tags should be used in a tagset goes back to early work on tagging corpora for computational purposes. The issues that guide and determine the design of a tagset may be purely linguistic or to the other extreme applicative. This distinction has strong connection to the method that is chosen to evaluate its quality. Tagging a corpus along with the development of a tagger may influence the tagset design in order to achieve better results and eliminate weak points of the tagger. The pioneering Brown Corpus was lexical-oriented, and its tagset was used as a baseline for many subsequent tagging projects. The Penn Treebank tagset was planned with a stochastic orientation and aimed to reduce redundancy, and therefore, elaborated the definition of tags to be less lexical and to carry less information that can be recovered automatically (e.g. past tense morphemes). In addition, tags with more general denotation are less bound to inconsistencies (e.g., compare a tagset with a single RB tag for all adverbs instead of tagset distinguishing RB and RN for nominal adverbs). The Penn Treebank tagging process was more syntactic and less lexical in nature, therefore, the same lexical item could be tagged differently in distinct syntactic contexts. In cases of disagreement among human annotators or where the POS was ambiguous, a word could be assigned more than one tag (Marcus et al., 1993).

Many tagging projects were influenced by English tagsets, which were used as the starting point for design for other languages as well. However, such tagset adoption is not a straightforward matter, and different language-families require careful treatment. Van Mol (2002) presents the problems of tagging the Arabic language, where words can be used in more than one syntactic function (an adjective used as noun), or even two lexical categories (both noun and adjective for the same lexeme). As in Hebrew, beinoni in Arabic can be used as adjectives, nouns, even prepositions and verbs. The proposed method tends towards the syntactic direction, allowing a word to be tagged according to its specific syntactic functions.

For the Hebrew Treebank project, the Penn Treebank tagset also served as a basis, however, due to the agglutinative and inflective morphological nature of Hebrew, complex tags (IN+PRP) were added and morphological features could be added to tags. The tagging approach of the Treebank was strictly syntactic, distinguishes for instance the tag CDT for numerals in determiner position and CD for other occurrences (Sima’an et al., 2001).

As mentioned above, a tagset design is influenced by the purpose of the tagging process, and therefore, there are various possible measures to test quality. Dejean (2000) distinguishes between internal (i.e., the quality of the tagger) and external measures. External quality; means *the extent to which it allows retrieval of all important grammatical distinction in the language* (Sampson cited by Dejean), practically – this was tested by evaluating how a tagset supports effective syntactic parsing.

4. Hebrew beinoni
As noted by the traditional Hebrew grammarian Gesenius (1976, p.355), the so-called beinoni form occupies a middle place between noun and verb. Morphologically, beinoni forms are simple nouns or adjectives, *i.e.*, they carry gender, number, and status inflections, prefixation, definiteness, and no person and tense/mood inflections. From the semantic point of view, according to traditional descriptions, Hebrew beinoni forms do not denote a fixed state, but activities, in contrast to nouns and adjectives.

There are many occurrences in the corpus where words in beinoni forms could not be assigned any of the traditional
| Word               | Example                                      | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|--------------------|----------------------------------------------|---|---|---|---|---|---|---|---|
| אهوּב               |数控 (LAH) אidges - אҳө́ב | N | N | A | N | A | N | N | N |
| 'ahub              | beloved                                      | V | A | A | A | N | N | N | N |
| Моск              | hadabar 'amur bmešine toqepp                 | A | A | V | A | A | A | V | A |
| 'אשֶמ               | 'ulay 'asem hamevudym hatelewisziomy         | A | V | A | N | N | N | N | N |
| 'אשֶמ               | maybe, the television medium is guilty       | N | N | A | A | A | A | A | V |
| בהלאָה            | belelah mehoser samkut                      | N | V | A | N | A | A | A | V |
| רודרֶבֶב          | hi' hadabar bimšutap 'al yedey kamah gupim   | A | A | A | A | A | A | A | N |
| במלשעַה          | she was guided by several groups together    | V | V | V | V | V | V | V | V |
| מָזוּקֵים         | yeš lid'og l'išun neged maziqim             | N | N | N | N | N | N | N | V |
| מָזוּקֵים         | smoking against pests should be applied      | A | A | A | V | V | V | V | V |
| הומַקָעֵרִים      | no'ada lišheh 'et hamukšarim minitel         | N | A | A | N | V | V | V | V |
| הומַקָעֵרִים      | intended to release the burden from the talented | V | V | V | V | V | V | V | V |
| הָנִמְנָה       | zeh lo’ min hanimna’                       | A | N | A | A | A | A | A | V |
| הָנִמְנָה       | it is possible that                         | V | V | V | V | V | V | V | V |
| מְשלָל             | hakoteb mšulal habanah taqtit              | A | A | A | A | A | V | V | V |
| מְשלָל             | the writer is bereft of any tactical knowledge | V | V | V | V | V | V | V | V |
| פּuestoָה        | hi’ sakbah pcu’ah qaše brošah             | N | A | A | N | N | N | N | V |
| פּuestoָה        | she was lying seriously wounded             | A | A | N | N | N | N | N | A |
| שָבֶה            | seper šobeh leb                            | N | V | V | V | V | V | V | V |
| שָבֶה            | an alluring book                           | V | V | V | V | V | V | V | V |
| ידוע              | yadu’ kih hakol hayah Şeber                | N | A | A | A | A | V | N | N |
| ידוע              | it is known that nothing was true           | V | V | V | V | V | V | V | V |
| ידוע              | yadu’ah bacibur                            | A | A | A | A | V | N | N | N |
| ידוע              | known in public                            | N | A | A | A | A | V | N | N |
| ידוע              | known                                    | N | A | A | A | A | V | N | N |

Table 1: Suggested POS for selected beinoni forms in various dictionaries.

tags, verb, noun or adjective. Consider the following example:

(1) גַּדוֹלִים (gdelim) (growing-up)?
    Morphologically, גדולם (gdelim) can be tagged as a masculine-plural adjective, or as participle inflection of the verb
    *לגדול (ligdol) (to grow). From a syntactic point of view, both
    these options are not possible: assuming this is a verb, the present tense cannot be substituted by future or past,
    without adding a covert relativizer
    *שינוויגדוֹלִים (bemsparim yelku wyigdlu) / *במספרים יכלו ויגדוֹלִים (in numbers that will grow / *in num-
    bers will grow). Assuming גדולם is an adjective, then coor-

2Transcription according to Ornan (2002).
dination requires the word holkim (going) to be an adjective as well, which is definitely not the case. Since we require categories to include only words that obey the same morphological and syntactic constraints, we conclude that in cases like this, these words must be assigned a distinctive tag.

We use the term beinoni to denote various forms of Hebrew tokens:

1. ‘Present verb like’ forms, with optional פדם prefixes, e.g., תפור, מגלגול, wešomrot (and are guarding/and guard/and guards).
   - holkim (dreaming/dreams)
   - samer (being guarded/is guarded)
   - šenišmarot (that are being guarded/those are guarded)
   - habolemet (that are dreaming/those dreams)
   - mšumar (is being conserved/is conserved/a conserve)
   - lanismarot (to (those that) are guarded)
   - kmšameret (as preserves/like a sifter)
   - mhamšumar (of the one that is conserved/of the conserve)
   - Ṣamur (at guards/at (those that) guards)

2. ‘Present verb like’ forms, with optional ב-imenti b,klm, e.g., ב-imenti bšomrim (at guards/at (those that) guards).
   - Ṣamur (at (those that) are guarded)
   - hastem (to (those that) are guarded)
   - lanišmarot (to (those that) are guarded)
   - Ṣomar (of the one that is conserved/over the conserve)
   - Ṣamur (at guards/at (those that) guards)

3. Construct state forms of nouns and adjectives, including those which are not part of the lexicon, e.g., שומרי šomrei (the guards/(those that) guard).
   - шомри šomre (the guards/(those that) guard)
   - שומרי šomrei (the guards/(those that) guard)
   - שומרי šomrei (the guards/(those that) guard)

4. Noun and adjective forms, including those which are not part of the lexicon, with pronominal suffix, שומרי šomrab (his guards/(those that) guard him).
   - שומרי šomrab (his guards/(those that) guard him)

There are four possible tags for these forms: present verb, participle, noun, and adjective. Each form may be tagged by different subsets of these tags:

1. Verb in present tense (with relativizer/subordinate conjunction), fixed noun or adjective – in case there is a lexicon entry for such noun or adjective – (with definite article), participle (with definite article or relativizer/subordinate conjunction).
2. Fixed noun or adjective (in case there is a lexicon entry for such noun/adjective), participle.
3. Noun or adjective in construct state (in case there is a lexicon entry for such noun/adjective), construct state of participle.
4. Noun or adjective with possessive pronoun (in case there is a lexicon entry for such noun/adjective), participle with accusative pronoun suffix.

5. A Lexical Category for beinoni

We are interested in the classification of the above beinoni forms. Rosen (1977, pp.106–107) argues for a participle category, which covers the participle and present verb forms. Blao (1966, p.186), on the other hand, treats the participle forms as either noun, adjective or verbs. A similar disagreement is found among modern analyzers: Rav Milim and Yona have no participle category, i.e., all the verbal interpretations are classified as verbs with a beinoni tense, which is the tense of the present forms as well, e.g., מגלול, שומרי (are guarding, are being rolled). Participles are classified in the lexicon into three categories: (1) ‘exclusively’ nouns/adjectives, with no possible verbal analysis, e.g., תמך, מתמחה, תמך, תמך, תמך, תמך (takuf, mlomad, soper (sewn, scholarly, writer), (2) nouns and adjectives, which have a verbal interpretation as well, e.g., מגלול, שומרי (rolled/is rolled, guard/guards, terrorists/sabotages), (3) exclusively verbal forms, e.g., אמר, הוורמאק, מɵר, מנוע (broadcast, counts, curses).

The Hebrew Knowledge Center morphological analyzer (http://mila.cs.technion.ac.il/) defines a participle category, generally composed of the beinoni tense verbs of Rav Milim and Yona.

This categorization decision is related to several issues: can the list of nouns and adjectives in the lexicon be extended by all participle forms? What is the correct reading of the ה prefix of the above beinoni forms: definite article or relativizer/subordinate conjunction (see Rosen (1977, pp.107, footnote 92))? Is there a conceptual difference between participle tense and present tense? Is there a hidden person mark for present and participle verbs? How do participles relate to generics formation in Hebrew?

In our analysis, we conclude that a distinct participle category should be defined. In contrast to the KC analyzer, we propose that present verbs should be assigned to the verb category, and distinguished from participles.

5.1. Morphology

From the morphological point of view, the beinoni forms are inflected and affixed as nouns, as shown in Table 2. According to Rav Milim, which has no beinoni category, the verb category contains tokens with status property, as

3 http://mila.cs.technion.ac.il/html

4 One might distinguish morphologically between nouns and adjectives by the existence of a pronominal suffix, which is somehow garbled for adjectives, e.g., tpuraw, ʻacubeyah (his sewns, her sadness), but we decided to consider such constructions as adjectives and not as nouns - see Adler (2007, section 5.3.2).
well as preposition prefixation. The KC analyzer, on the other hand, combines participles and present verbs, which have a different affixation mechanism and status marking, under the same *participles* category.

5.2. Syntax

From a syntactic point of view, certain noun/adjective *beinoni* forms, cannot be considered as verbs nor as nouns/adjectives.

5.2.1. Noun/adjective usages that cannot be considered as verbs

**Tense Affinity**

Noun/adjective usages have no tense affinity, in contrast to present verbs (Blao, 1966, p. 186). The same surface form (*beinoni*) can be used as a noun/adjective or as a present verb. How can we distinguish between these two usage types? Present verb usages are bound to present tense, while noun/adjective can occur in any tense context. Aspect is not relevant to this distinction – *beinoni* in verbal usage can denote both progressive and simple tenses (in contrast to the English present participle which is bound to the progressive aspect).

The following examples indicate simple syntactic tests that distinguish between verbal and noun/adjective usages:

(2) "החיילים מתאמנים" (verbal usage: present progressive) "the-soldiers train now"

\[ * \text{החיילים מתאמנים אתמול} \]

*"the-soldiers train yesterday"

(3) "החיילים מתאמנים בימים אלו" (verbal usage: present simple) "the-soldiers train these days"

\[ * \text{החיילים מתאמנים בימים מהם} \]

*"the-soldiers train those days"

5.2.2. Noun usages that cannot be considered as nouns

**Complement**

A complement is not necessarily required for nouns in contrast to noun usages of *beinoni* form of transitive verbs (Shlonsky, 1997, pp. 27–28).

(5) "היא מנהיגה את הקבוצה" [verb] "she is-leading ACC the-group"

(6) "היא מנהיגה" [noun] "she is a leader"

(7) "הוא לוכד נשים" [verb] "he traps/is-trapping snakes"

(8) *"הוא לוכד" [beinoni] *he traps/is trapping"

(9) *"הוא לוכד של ministra החקלאות" [beinoni, possessive pronoun] *he traps POSS ministry the-agriculture"

Shlonsky (1997, chapters 2-5) claims that verbal *beinoni* is a participle, and Hebrew has a null auxiliary. Shlonsky employs Chomsky’s *government and binding* approach in order to present tense sentences on a par with compound tense constructions – *beinoni* is a hybrid form, a verb whose agreement features are participial but raised to T. In spite of his elegant word-order and clause-structure analysis, we prefer, for our purpose, to avoid modeling syntactic movements, and formalize a definition which is based on the tokens as they appear in the text.

Explicit Subject

Noun/adjective usages do not require an explicit subject. *Beinoni* in verbal usages require an explicit subject, which can be absent from noun/adjective constructions, i.e., the token מטפסים המטפסיםوشしまう (climbers were given water) can only be interpreted as a noun/verb usage but not as a present verb.

5.2.2. Noun usages that cannot be considered as nouns

**Genitive של**

Noun usage of *beinoni* cannot modify the genitive ל של or be suffixed by a possessive pronoun, in contrast to regular nouns (Shlonsky, 1997, pp. 27–28).

(9) *"הוא מנהיגת של קבוצת" [noun] *she is a-leader POSS a-group"

(10) *"הוא מנהיג" [noun] *she is their leader"

(11) *"הוא לוכד של ministra החקלאות" [beinoni, possessive pronoun] *he traps POSS ministry the-agriculture"

(12) *"הוא לוכד" [beinoni] *he traps-POSS"

3171
Table 2: Morphological classification of participle forms.

| Gender | Number | Status | w, h | b, k, l, m | suffix |
|--------|--------|--------|------|------------|--------|
| Noun   | V      | V      | V    | V          | V      |
| Adjective | V | V      | V    | V          | X      |
| Present Verb | V | V      | X    | V          | X      |
| Beinoni | V      | V      | V    | V          | V      |

On the other hand, accusative pronoun suffix, and/or accusative modification by a preposition של šel, is possible for noun usage of beinoni form.

(13) מחלד [beinoni, accusative pronoun] halten
he traps them

(14) מחלד של נשים [beinoni, accusative pronoun] halten šel nhašim
he is a snake trapper

**Construct state** Construct state of regular nouns can be either possessive (as nouns) or accusative (as beinoni), in contrast to construct state of benoni which is always accusative.

(15) שמיר [noun]
šomrei hamip’alim
guards factories

(16) חלד ידוע [beinoni]
lokei hanhašim
trap snakes

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

(17) השומר של המפעלים
šomrer šel hamip’alim
guarding the factories

(18) כל השומר של המפעלים ידוע את תפקידו
kol šomrer šel hamip’alim yode’ ‘et tapqido
all the guards knows his duty

Adjective modifier An adjective modifier is possible for noun usgae of beinoni, in contrast to present verb (Shlonsky, 1997, pp. 27–28).

(19) מנהיגה דגולה [noun]
hi’ manhigah dgulah
she is a leader outstanding

5.2.3. Adjective usages that cannot be considered as adjectives
Certain adjective usages of beinoni forms do not stand for the adjective tests, suggested by Doron (2000).

**Negation** The negation prefix בולתי bilti modifies adjective, in contrast to adjective usage of beinoni.

(20) ידוע מהסומך bilti musmak
un certified
6. Our Guidelines

In our final version of the tagging guidelines, four different POS tags can be proposed for the various forms of beinoni, by the morphological analyzer. The tagger must select among the possible tags based on the context:

- **Noun** – should be suggested by the analyzer for any form which is listed in the lexicon as a noun. The noun list should be extended by any beinoni form of the verbs in the lexicon, if the corpus contains instances of these forms in a role according to lexicographic noun phrase construction tests (listed in Adler (2007, Appendix B.1.1)).

- **Adjective** – should be suggested for any form which is listed as an adjective in the lexicon. The adjective list should be extended by any beinoni form of the verbs in the lexicon, if the corpus contains instances of these forms in a role of adjective according to lexicographic adjective phrase construction tests (listed in Adler (2007, Appendix B.1.2)).

- **Participle** – the participle option should be suggested only for absolute state forms, which have no suffix or מ ל כ ב k l m prefixes.

- **Verb** – a present-tense verb analysis should be suggested only for absolute state forms, which have no suffix or מ ל כ ב k l m prefixes.

7. Experiments

With these guidelines, an agreement of above 99% was reached among 4 human taggers with respect to the definition of participle, verb, noun, and adjective categories. The ambiguity level of the analyzer, i.e., the average number of analyses per token, was not significantly changed. Following Dejean (2000), we use Hebrew Simple NP chunking (Goldberg et al., 2006) as an external application on which to test our tagset. Chunking NPs is advantageous for this task as participles and NPs are closely related. Our chunks definition is based on that of Goldberg et al. (2006), with the exception that chunk boundaries are not allowed to break orthographic token boundaries. We trained 3 SVM-based chunking models (Goldberg et al., 2006; Kudo and Matsumoto, 2000), each with a different tagset on the same data. We used the same feature set and SVM configuration for all models.

The tagsets we used were: (1) all beinoni forms are tagged as Participle (Part), (2) beinoni forms are tagged as either Participle or Present Verb (Part + V) and (3) each beinoni form is tagged as either Noun, Verb or Adjective (NoPart).

Looking at the train and test sets, the beinoni forms appear in less than 3% of the tokens. Of the beinoni forms, 90% are present-Verbs. This leaves about 50 non-Verbal beinoni forms and 297 Verbal beinoni forms in the test corpus. The resulting chunk accuracies (F) were: 91.09 (Part), 91.23 (NoPart), 91.31 (Part + V). While the number

---

5We looked at a 5 morphemes window surrounding the word to be classified, and considered the lexical form, POS and construct-state information as features. We used a polynomial kernel of degree 2.
are very close to each other, our proposed tagset performs marginally better for this task. Considering the low count of beinoni occurrences in the corpus, and the even lower count of non-verbal beinoni forms, one can not expect to achieve bigger improvements on an external task. The experiment verifies that distinguishing present-verbs from participles improves chunking, and that noun and adjectival uses of the beinoni form should be grouped together. The experiment verifies that our proposed tagset does not harm chunking performance, while being linguistically justified and greatly improving the agreement between human annotators.

8. Conclusion

This paper illustrates the issues faced when designing a tagset for POS tagging for Hebrew. Our objectives are to ensure high consistency among human taggers, to offer adequate linguistic description, and to verify that the tagset allows us to perform precise machine learning for syntactic parsing. We specifically investigated the decision to introduce a distinct tag for beinoni forms in Hebrew. We have verified that with proper guidelines, and an adapted lexicon, this participle tag allowed us to reduce inconsistent manual tagging errors (increased internal tagging quality). From the linguistic point of view, we justify the addition of new lexical category - participle. In contrast to Rosen (1977) and the KC analyzer, our new category excludes present verbs.

Although evaluation on chunking did not show significant improvement, it does verify that the addition of such category does not harm chunking. Besides linguistics arguments, practical considerations (e.g., the easiness of tagging, agreement among taggers and dictionaries) strongly support the usage of this category in the Hebrew tagset.

9. Acknowledgements

This work is supported in part by the Lynn and William Frankel Center for Computer Sciences.

10. References

Meni Adler and Michael Elhadad. 2006. An unsupervised morpheme-based HMM for Hebrew morphological disambiguation. In Proceeding of COLING-ACL-06, Sydney, Australia.

Meni Adler. 2007. Hebrew Morphological Disambiguation: An Unsupervised Stochastic Word-based Approach. Ph.D. thesis, Ben-Gurion University of the Negev, Beer-Sheva, Israel.

Eitan Avneyon, Raphael Nir, and Idit Yosef. 2002. Milon sapir: The Encyclopedic Sapphire Dictionary. Hed Artsi, Tel-Aviv, Israel. (in Hebrew).

Yehoshua Biao. 1966. Syntax Fundamentals. Hebrew Institute for Written Education, Jerusalem. in Hebrew.

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

Yaacov Chooueka, Uzi Freidkin, Hayim A. Hakohen, , and Yael Zachi-Yannay. 1997. Rav Milim: A Comprehensive Dictionary of Modern Hebrew. Steimatsky, Tel-Aviv, Israel. (in Hebrew).

Herve DeJean. 2000. How to evaluate and compare tagsets? a proposal. In Proceedings of LREC 2000, Athens, Greece.

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

Avraham Even-Shoshan. 2003. Even Shoshan’s Dictionary - Renewed and Updated for the 2000s. Am Oved, Kineret, Zmora-Bitan, Dvir and Yediot Aharonot. (in Hebrew).

Friedrich H. W. Gesenius. 1976. Hebrew Grammar. The Clarendon Press, Oxford. Edited and enlarged by E. Kautzsch, English edition by A. E. Cowley. Yoav Goldberg, Michael Elhadad, and Meni Adler. 2006. Noun phrase chunking in hebrew influence of lexical and morphological features. In Proceeding of COLING-ACL-06, Sydney, Australia.

Yaakov Knaani. 1960. The Hebrew Language Lexicon. Masada, Jerusalem, Israel. (in Hebrew).

Taku Kudo and Yuji Matsumoto. 2000. Use of support vector learning for chunk identification. In Proceedings of CoNLL-00 and LLL-00, Lisbon, Portugal.

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

Mark Van Mol. 2002. The semi-automatic tagging of arabic corpora. In Arabic Language Resources and Evaluation - Status and Prospects Workshop, LREC.

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

Uzi Ornan. 2002. Hebrew in Latin script. Léshonénu, LXIV:137–151. (in Hebrew).

Haim B. Rosen. 1977. Contemporary Hebrew. Mouton, The Hague, Paris.

Beatrice Santorini. 1995. Part-of-speech tagging guidelines for the Penn Treebank Project. 3rd revision:. Technical report, Department of Computer and Information Science, University of Pennsylvania.

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

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

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

Shlomo Yona. 2004. A finite-state based morphological analyzer for Hebrew. Master’s thesis, Haifa University.