Tagging a Hebrew Corpus: The Case of Participles

Meni Adler and **Yael Netzer** and **Yoav Goldberg** and **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 4 aspects: the tagset should be consistent with common linguistic knowledge; there should be maximal agreement among taggers as to the tags assigned to maintain consistency; the tagset should be useful for machine taggers and learning algorithms; 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 participles in Hebrew. We explain how this tag is defined, and how it helped us improve the 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 participles in Hebrew. The issue is how participles should be tagged. Three main approaches are considered: treat participles as either verbs, nouns and adjectives according to the context; treat participles as verbs; or – the approach we adopt – adding 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 participles, 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 our-

selves 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 fulfil 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 and uninformative tagset of one tag (WORD) would be optimal. Most meaning-carrying words belong to one of the main three categories. 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. One of the main confusing factor we found among taggers was related to the status of what we call "participles". 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 roughly 27M tokens, in articles of typical length between 200 to 1000 tokens. Of the full corpus, a sample of articles comprising altogether 200K tokens was assembled at random, This sample was manually tagged for part of speech (for more 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. Given the lack of a reliable lexicon, the taggers were eventually 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 to 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 participles. We focus in this paper on participles.

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 participles, as determined by: (1) Rav Milim [Choueka et al., 1997], (2) Sapir [Avneyon et al., 2002], (3) Even-Shoshan [Even-Shoshan, 2003], (4) Knaani [Knaani, 1960], (5) HMA [Carmel and Maarek, 1999], (6) Segal [Segal, 2000], (7) Yona [Yona, 2004], (8) 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-level 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 stochastic orientation and aimed to reduce redundancy, and therefore, elaborated the definition of tags to be less

Word	Example	1	2	3	4	5	6	7	8
אהוב ***’ahub*** **beloved**	זר דפנים לגיבור - אהוב *zer dpanim lgibor - ’ahub* a garland of laurels for a beloved hero	N V	N A	A	N A	A N	N V	N	N
אמור ***’amur*** **shouldn**	הדבר אמור במשנה תוקף *hadabar ’amur bmišne toqep* It is said with strength	A	A	V	A	A	V X	A	V
אשם ***’ašem*** **guilty**	אולי אשם המדיום הטלוויזיוני *’ulay ’ašem hamedyum haṭelewizyoni* maybe, the televisional medium is guilty	A	V	A	A	N A	N A	N A	N A V
בטלה ***btelah*** **is cancelled**	בטלה מחוסר סמכות *btelah meḥoser samkut* is cancelled due to lack of authorization	N A	V	A V	A	N A	N A	N A V	V
במשותף ***bimšutap*** **in common**	היא הודרכה במשותף על ידי כמה גופים *hi’ hudrkah bimšutap ‘al yedey kamah gupim* she was guided by several groups together	A V	A V	A V	A V	A V	A	A	N
ישוב ***yašub*** **seated**	ישוב באחוזתו היפהפיה *yašub b’aḥuzto haypeypiyah* seated in his lovely estate	A	A	A	A R	A	V	A	V
מזיקים ***maziqim*** **pests**	יש לדאוג לעישון נגד מזיקים *yeš lid’og l‘išun neged maziqim* smoking against pests should be applied	N A V	N A V	N V	N A V	N A V	N A	V	A
המוכשרים ***hamukšarim*** **the talented**	נועדו לשחרר את המוכשרים מנטל *no‘adu lšaḥrer ’et hamukšarim minetel* intended to release the burden from the talented	A V	A V	A V	N A V	V	N V	A V	N
הנמנע ***hanimna‘*** **avoidable**	זה לא מן הנמנע *zeh lo‘ min hanimna‘* it is possible that	A V	N A V	A V	N A V	A V	A V	V	N
משולל ***mšulal*** **bereft**	הכותב משולל הבנה טקטית *hakoteb mšulal habanah ṭaqṭit* the writer is bereft of any tactical knowledge	A	A	A V	A V	A V	A	N	A
פצועה ***pcu‘ah*** **wounded**	היא שכבה פצועה קשה בראשה *hi’ šakbah pcu‘ah qaše brošah* she was lying seriously wounded	N A	A	A V	N A	N A	N V	N A	V
שובה ***šobeh*** **captures**	ספר שובה לב *seper šobeh leb* an alluring book	N V	V	V	N V	V	N	N V	A
ידוע ***yadu‘*** **known**	ידוע כי הכל היה שקר *yadu‘ ki hakol hayah šeqer* it is known that nothing was true --- ידועה בציבור *yadu‘ah bacibur* known in public	N A	A	A	A	A	A V	N A	N V

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

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. [Mol, 2002] presents the problematics 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, participles 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 – parsing.

4 Hebrew Participles

As noted by the traditional Hebrew grammarian Gesenius [Gesenius, 1976, *p*.355], the so-called *beinoni* form (which we translate 'participle') occupies a middle place between noun and verb. Morphologically, participles 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 participles do not denote a fixed state, but activities, in contrast to nouns and adjectives.

There are many occurrences in the corpus where words in participle forms could not be assigned any of the traditional tags, verb, noun or adjective. Consider the following example:

(1) היום נכונים ישראלים, במספרים גדלים והולכים, לקלוט את הסיסמא
hayom nkonim yiṣr'elim, bmisparim gdelim wholkim, liqloṭ 'et hasisma'
today ready Israelis, in-numbers growing and-going, to-accept the-slogan.
Nowadays, Israelis are ready and willing, in growing numbers, to accept the slogan.

How can we tag the word גדלים? Morphologically, גדלים 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 option are not possible: assuming this is a verb, the present tense cannot be substituted by future or past, without adding a covert relativizer (במספרים שילכו ויגדלו)(*במספרים ילכו ויגדלו). Assuming גדלים is an adjective, then coordination requires the word הולכים 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 ו,ש,ה *w,š,h* prefixes, *e.g.*, ושומרות *wešomrot* (and are guarding/and guard/and guards).

2. 'Present verb like' forms, with ב,כ,ל,מ *b,k,l,m*,

e.g., בשומרים *bšomrim* (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).

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

We are interested in the classification of these *beinoni* forms. Rosen [Rosen, 1977, *pp*.106–107] argues for a *participle* category, which covers the participle and present verb forms. Blao [Blao, 1966, *p*.186], on the other hand, treats the participle forms, as either noun or adjective, as 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.*, שומרים, מגולגלים *šomrim, mgulgalim* (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.*, תפור, מלומד, סופר *tapur, mlumad, soper* (sewn, scholarly, writer), (2) nouns and adjectives, which have a verbal interpretation as well, *e.g.*, מגולגל, שומר, מחבל *mgulgal, šomer* (rolled/is rolled, guard/guards, terrorists/sabotages), (3) exclusively verbal forms, *e.g.*, משודר, סופר, מחרף *is broadcast, counts, curses*. The Hebrew Knowledge Center morphological analyzer (KC) [Adler, 2007, section 4.2] defines a *participle* category, generally composed of the *beinoni* tense verbs of Rav Milim and Yona.

This matter 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 *h* 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.

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 participle form of the verbs in the lexicon, if the corpus contains instances of these forms in a noun role according to lexicographic noun phrase construction tests.

- Adjective – should be suggested for any form which is listed as an adjective in the lexicon. The adjective list should be extended by any participle 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.

- Participle – the participle option should be suggested for any of the *beinoni* forms.

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

5 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.

Following Dejean [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 ortographic token boundaries. We trained 4 SVM-Based chunking models [Goldberg

et al., 2006; Kudo and Matsumato, 2000], each with a different tagset on the same data. We used the same feature set and SVM configuration for all models. The 4 tagsets used were: the original Hebrew Treebank POS tags (TB), the Treebank POS tags in which instances of the MOD tag were manually resolved to either NN/JJ/RB ($TBConv$), our tagset without present verbs (ANN) and our tagset with present verbs ($ANNV$). In the SVM, we considered only POS and construct-state information as features.

The resulting chunk accuracies (F) were: 94.01 (TB), 93.89 ($TBConv$), 91.28 (ANN), 91.31 ($ANNV$). It seem that the TB based tagset is superior to our tagset for this task. However, error analysis reveals that the main cause of lower accuracy for the non-treebank tagset is that the chunks were automatically derived from the Treebank. Therefore, chunk-boundaries and TB POS tagging decisions are closely tied together – any inconsistencies in the Treebank were reflected on both the POS tags *and* the chunk boundaries. Moreover, individual tagging decisions in the treebank are reflected in the chunk structure, and vice-versa (*e.g.*, choosing to tag a certain participle as a VB instead of a JJ induces a chunking decision in which this participle is not included in the chunk). Our on tagging, on the other hand, is consistent within itself but does not consider the treebank chunk-boundaries. That is, our tagset might be inferior for the task of identifying the specific chunks derived from the Treebank, but not on the general task of identifying syntactic chunks. The fact that the $TBConv$ results are very close to those of TB supports this claim – it is not so much the specific tags that make the difference, but their association to the chunk boundaries. We are currently in the process of manually annotating chunk boundaries on the same data set, in order to verify this claim.

6 Conclusions

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 participles 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), while still ensuring accurate machine learning for a syntactic chunking task (external tagging quality).

References

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

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

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

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

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

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

[Dejean, 2000] Herve Dejean. 2000. How to evaluate and compare tagsets? a proposal. In *Proceedings of LREC 2000*, Athens, Greece.

[Even-Shoshan, 2003] 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).

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

[Goldberg et al., 2006] 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.

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

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

[Marcus et al., 1993] 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.

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

[Netzer et al., 2007] Yael Netzer, Meni Adler, David Gabay, and Michael ELhadad. 2007. Can you tag the modal? you should! In *Proceeding of COLING-ACL-07*, Prague, Czech.

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

[Santorini, 1995] 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.

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

[Sima'an et al., 2001] 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.

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