Tagging a Hebrew Corpus: The Case of Participles

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

Department of Computer Science | **Ben Gurion University of the Negev**

What tagset should be used for Hebrew?
Adopting an English tagset does not work. A Hebrew-specific tagset must be designed.

The Hebrew *Beinoni* Form

Roughly correspond to *Participles*:

ראיתי חמישה רוכבים
see-1sg-past five ride-participle
I saw five riders

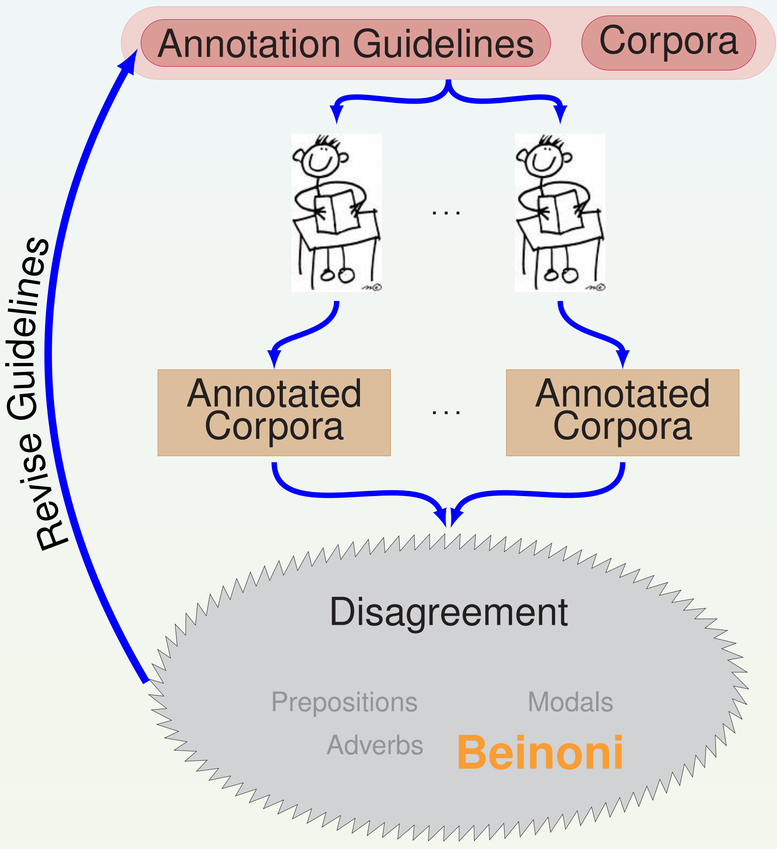
Beinoni Forms are Hard to Tag

Nouns? Verbs? Adjectives?

- ∼70% disagreement among 4 annotators

Our Goals

- Better linguistic characterization of the *Beinoni* forms
- Better tagging guidelines
  ⇒ Better annotators agreement



Beinoni in Traditional Dictionaries

No agreement among dictionaries on the part of speech given to *Beinoni* forms.

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 hatelewizyoni* maybe, the television 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'ahuzto 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
המכשירים *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 taqtit* 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

Linguistic Distinctions

Morpho-Lexical Classification of Beinoni Forms

Morphologically, *Beinoni* has the form of present verb, with prepositional בכלמ prefixation, and construct state inflection.

We distinguish:

Unlexicalized-Beinoni form that is not lexicalized as a noun or an adjective in the lexicon, *e.g.*, לכוד.

Verbal-Beinoni 'present' like *Beinoni* form (absolute state, with no בכלמ prefixes).

Nominal-Beinoni form that is lexicalized as a noun, *e.g.*, שומר ("guarding", but also "a guard").

Adjectival-Beinoni form that is lexicalized as an adjective, *e.g.*, סגור ("closed").

Not all *Beinoni* Forms are Nouns

Beinoni can enter in the same syntactic constructions as nouns - but there are systematic differences.

Nouns don't require a complement, but transitive verb Beinoni do:

הוא לוכד נחשים / * הוא לוכד
he traps/is-trapping snakes
*he traps/is-trapping

Nouns can be modified by the של genitive, but *Beinoni* cannot:

הוא לוכד של משרד החקלאות *
*he traps POSS ministry the-agriculture
*he traps of the agriculture ministry

Pronominal suffix on Nouns indicate possessiveness. A pronominal suffix on *Beinoni* is accusative.

הוא לוכדם *
*he traps-POSS → he traps-ACC
*he traps of them → he traps them

The construct state of nominal-beinoni is either possessive or accusative, but for *Beinoni* forms, the construct state is always accusative.

לוכדי הנחשים
⇒ הלוכדים של הנחשים*
⇒ הלוכדים את הנחשים
the snake trappings-CONST
⇒ *the trappers belonging to snakes
⇒ the snake trappers

The prefix ה represents a definite article for regular nouns, and a relativizer for unlexicalized *Beinoni* forms.

Not all *Beinoni* Forms are Present Verbs

Morphologically

Verbs cannot be prefixed by prepositions (בכלמ)

פניתי לרוכבים / I talked *to* riders

Verbs don't have a construct form

מטפסי הרים / mountain climbers

Syntactically

Present verbs are bound to present tense...

הילדים מתאמנים עכשיו / * אתמול
Kids are training now / *yesterday

...while *Beinoni* can occur in any tense context.

המתאמנים הגיעו / The training arrived.
The trainees arrived

In addition, present verbs *require explicit subject*, *cannot function as subjects* on their own, and *cannot be quantified*, in contrast to *Beinoni*:

שני מתאמנים הגיעו
two training arrived
Two trainees arrived

Not all *Beinoni* Forms are Adjectives

Unlexicalized-Beinoni cannot be negated by the prefix בלתי.

בלתי בטל / * בלתי מוסמך
un-certified / *un-insignificant

Unlexicalized-Beinoni cannot appear as complements of the verbs נותר (*remain*) and נראה (*seem*).

הוא נותר בטל / * הוא נותר עייף
he remains tired / *he remains insignificant

Tagset Design

A (new) Tag in Hebrew: `Beinoni`

- Any *Beinoni* form can be tagged as `Beinoni`
- A *Beinoni* form can be tagged as `Noun`/`Adjective` when lexicalized
- A *Beinoni* form can be tagged as `Verb` only if there is no בכלמ preposition attached

Evaluation

Internal Evaluation: Tagging Agreement

Using refined tagset and guidelines, 4 annotators reach **99%** agreement on *Beinoni* forms.

Tagset Performance on an *External Task*: SVM Based NP Chunking

- No `Beinoni` tag ⇒ 91.23F
- All *Beinoni* forms are tagged as `Beinoni` ⇒ 91.09F
- **Our Proposed Tagset/Guidelines** ⇒ 91.31F

The proposed tagset does not hurt and slightly facilitates an external task.

Selected References

- Meni Adler. 2007. *Hebrew Morphological Disambiguation: An Unsupervised Stochastic Word-based Approach*. Ph.D. thesis, Ben-Gurion University of the Negev, Beer-Sheva, Israel.
- 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.
- Edit Doron. 2000. The passive participle. *Hebrew Linguistics*, 47:39–62. (in Hebrew).
- Ur Shlonsky. 1997. *Clause Structure and Word Order in Hebrew and Arabic*. Oxford Universirt Press, New York Oxford.
- Haim B. Rosen. 1977. *Contemporary Hebrew*. Mouton, The Hague, Paris.
- Yehoshua Blao. 1966. *Syntax Fundamentals*. Hebrew Institute for Written Education, Jerusalem. in Hebrew.
- 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.