# Word-Based or Morpheme-Based? Annotation Strategies for Modern Hebrew Clitics

**Reut Tsarfaty** | Institute for Logic, Language and Computation | University of Amsterdam
Plantage Muidergracht 24, 1018TV Amsterdam, Netherlands | rtsarfat@science.uva.nl

**Yoav Goldberg** | Computer Science Department | Ben Gurion University of the Negev
P.O.B 653 Be'er Sheva 84105, Israel | yoavg@cs.bgu.ac.il

## Introduction

### Background
In morphologically rich languages a single word may carry different sorts of information and the different morphs may indicate a relation to other elements in the syntactic tree.

### The Question
Should we analyze a word as a sequence of morphological units or should we treat orthographic (space-delimited) words as the primitive units of our analyses?

### Our Investigation
We discuss and evaluate the adequacy of Morpheme-based and Word-based annotation strategies for the development of statistical parsers for Modern Hebrew.

**Morpheme-based (MB)** theories (Bloomfield, 1933; Hocket, 1954) assume that the atomic units of the language are morphs which are combined to create words through various processes (Matthews, 1991).

**Word-Based (WB)** theories consider words the atomic units of the language, and morphological considerations predict generalizations concerning the syntactic behavior of morphologically similar words (Blevins 2006).

## The Data

### Pronominal Clitics in Modern Hebrew

**(1) Prepositions** — *hwa ba ali* | הוא בא אלי

a. *hwa ba al ani*
he came to to I
He came to me

b. *hwa ba ali*
he came to.1p.sing
He came to me

**(2) Possessive Markers** — *hildim flnw* | הילדים שלנו

a. *hildim fl anxnw*
the-children of we
Our children

b. *hildim flnw*
the-children of.1p.masc.plural
Our children

**(3) Accusative Markers** — *hwa rah awth* | הוא ראה אותה

a. *hwa rah at hia*
he saw ACC she
He saw her

b. *hwa rah awth*
he saw ACC.3p.fem.sing
He saw her

**(4) The Dative Shift in Modern Hebrew**

a. *ntti lw mtnh* | נתתי לו מתנה
gave.1p.sing to.3p.masc.sing a-present
I gave him a present

b. **ntti mtnh lw* | *נתתי מתנה לו
*gave.1p.sing a-present to.3p.masc.sing
*I gave a present to him

c. *ntti lild mtnh* | נתתי לילד מתנה
gave.1p.sing to-the-child a-present
I gave the child a present

d. *ntti mtnh lild* | נתתי מתנה לילד
gave.1p.sing a-present to-the-child
I gave a present to the child

**(5) Coordinated Structures in Modern Hebrew**

a. *liwab wrewt* | ליואב ורעות
for-Yoav and-Reut
for Yoav and Reut

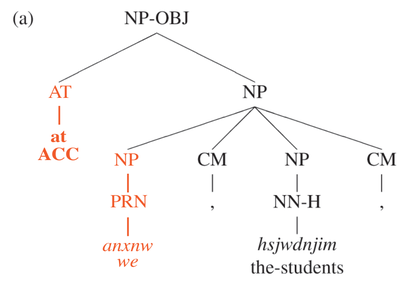
b. **li wrewt* | *לי ורעות
*for.1p.sing and-Reut
*for me and Reut

c. *li wliwab* | לי וליואב
for.1p.sing and-for-Yoav
for me and for Yoav

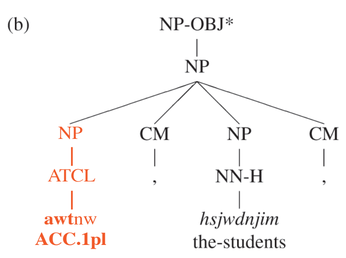
d. *li wlw* | *לי ולו
for.1p.sing and-for.3p.masc.sing
for me and for him

## Annotation Strategies

### A Morpheme-Based Strategy
The original annotation strategy in the Modern Hebrew Treebank.
**Advantage** | Capture correctly word- and constituent-boundaries discrepencies.
**Disadvantage** | Does not correspond deterministically to surface forms.

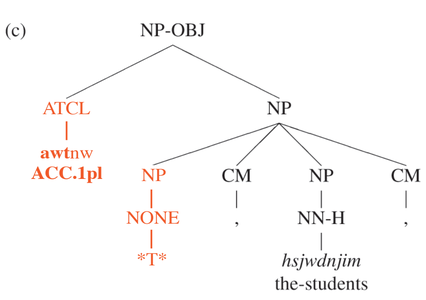



### A Naïve Word-Based Strategy
A naïve proposal for representing words as the yields of syntactic parse trees.
**Advantage** | Yields correspond directly to surface forms.
**Disadvantage** | grammatically incorrect.




### Our Word-Based Strategy
We propose an alternative Word-Based (WB) analysis as inflectional features on top of specialized categories of prepositions / markers. The special categories capture membership in distinct syntactic classes, and the features indicate agreement with a dropped pronoun marked as a trace of an empty element.

**Advantages** | recover correctly word- and constituent-boundaries discrepancies, represent yields directly as surface forms.




אותנו, הסטודנטים,
awtnw, hsjwdnjim,
ACC.1p.plural, the-students,
us, the students,

## Experimental Design

**Goal** | Compare and contrast the adequacy of the MB and WB annotation strategies for Modern Hebrew pronominal clitics

**Methodology** | Evaluate parsing performance of different treebank PCFGs corresponding to different annotation strategies.

**Data** | The Modern Hebrew Treebank version 1.0 (Sima'an et al., 2001), 5000 sentences from the daily newspaper `Ha'aretz'. The Test-Set constitutes the first 500 non-empty sentences.

**Grammar** | Extracted from tree-skeletons in which syntactic categories are extended with a handful of morphosyntactic features (marking infinitivals, definite phrases, Prepositional Possessive Phrases)

**Evaluation** | Problem: Different strategies result in different sentence length
Quantitative: compare PARSEVAL measures of sentences without clitics
Qualitative: analyze differences in the parse-trees that include clitics

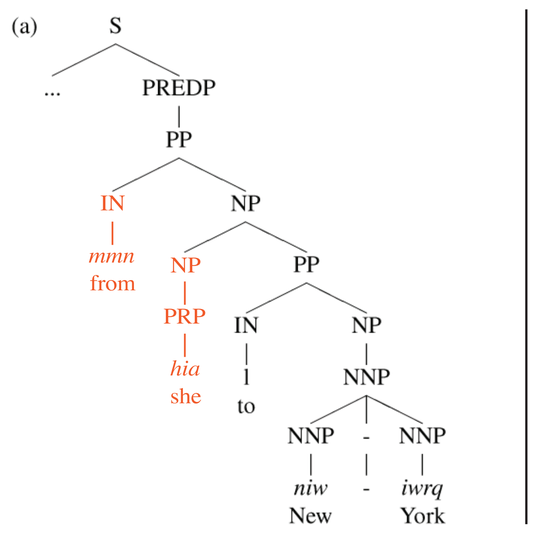
## Results and Analysis

### Quantitative Analysis
For sentences without pronominal clitics the conversion does not have a significant influence on the disambiguation capacity of the resulting grammars.

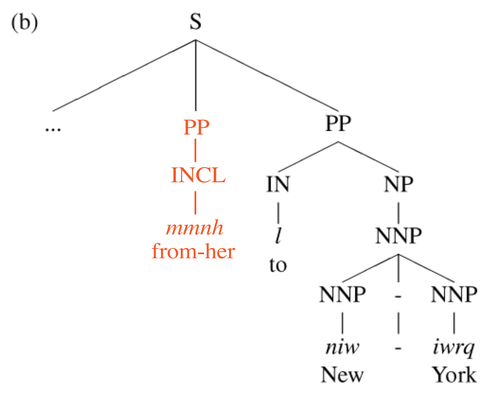
| | *No Clitics* before | *No Clitics* after | *Bare Prepositions* before | *Bare Prepositions* after |
|---|---|---|---|---|
| Prepositions (WP) | 79.15/80.94 (80.03) | 79.03/80.87 79.94 | 78.95/80.91 79.92 | 78.84/80.83 79.82 |
| Prepositions (WOP) | 80.57/82.42 (81.48) | 80.45/82.35 81.39 | 80.28/82.30 81.28 | 80.17/82.22 81.18 |
| Possessive Markers (WP) | 78.51/80.27 (79.38) | 78.52/80.28 (79.39) | 76.02/77.68 (76.84) | 76.02/77.73 (76.86) |
| Possessive Markers (WOP) | 79.89/81.72 (80.79) | 79.90/81.73 (80.80) | 77.51/79.26 (78.38) | 77.51/79.31 (78.40) |
| Accusatives Markers (WP) | 78.68/80.57 (79.61) | 78.63/80.55 (79.58) | 77.55/79.89 (78.70) | 77.39/79.83 (78.59) |
| Accusatives Markers (WOP) | 80.00/81.97 (80.97) | 79.97/81.96 (80.95) | 78.95/81.40 (80.16) | 78.83/81.37 (80.08) |

### Qualitative Analysis
For sentences with pronominal clitics the WB analysis is always as good or better than the original MB analysis.

| Parsing Result | Num of Sentences |
|---|---|
| Identical Parses | 116 |
| Only WB Correct | 5 |
| Only MB Correct | 0 |
| Both Wrong, WB Better | 8 |
| Both Wrong, MB Better | 0 |
| Both Wrong, None Better | 2 |




MB: Incorrect PP attachment




WB: correct PP attachment

The main source of errors for the MB strategy is its tendency to learn high attachment for prepositions that originate from cliticized elements.

## Conclusion

The WB analyses are more faithful to the surface forms thus avoiding the need for preceding segmentation.

The WB resulting treebank grammars provide better PP attachment disambiguation capacity.

The **Word-Based (WB)** annotation strategy is more adequate than the **Morpheme-Based (MB)** strategy for training statistical parsers on the Modern Hebrew Treebank.

## References

J. P. Blevins. 2006. Word-Based Morphology. Journal of Linguistics, 3(42):531-573. L. Bloomeld. 1933. Language. University of Chicago Press.

S. Cohen and N. Smith. 2007. Joint Morphological and Syntactic Disambiguation. In Proceedings of EMNLP-CONLL.

Y. Goldberg and R. Tsarfaty. 2008. A Single Generative Probabilistic Model for Joint Morphological Segmentation and Syntactic Parsing. In Proceedings of ACL.

C. F. Hocket. 1954. Two Models of Grammatical Description. Word, (10).

P. H. Matthews. 1991. Morphology. Cambridge University Press.

H. Schmid. 2004. Efficient Parsing of Highly Ambiguous Context-Free Grammars with Bit Vectors. In Proceedings of ACL.

K. Sima'an, A. Itai, Y. Winter, A. Altman, and N. Nativ. 2001. Building a Tree-Bank of Modern Hebrew Text. In Traitement Automatique des Langues.

R. Tsarfaty. 2006. Integrated Morphological and Syntactic Disambiguation for Modern Hebrew. In Proceeding of COLING-ACL SRW.

A. Zwicky. 1977. On Clitics. Indiana University Linguistics Club.

## Acknowledgments

We thank Meni Adler, James P. Blevins, Michael Elhadad, Khalil Sima'an, Remko Scha and Jelle Zuidema for comments and discussion.
We acknowledge the Netherlands Organization for Scientific Research (NWO) and the Lynn and William Frankel center for Computer Sciences in support of the work of R. Tsarfaty and Y. Goldberg respectively.