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

Reut Tsarfaty **Yoav Goldberg**

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

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

Abstract

Morphologically rich languages pose a challenge to the annotators of treebanks with respect to the status of orthographic (space-delimited) words in the syntactic parse trees. In such languages an orthographic word may carry various, distinct, sorts of information and the question arises whether we should represent such words as a sequence of their constituent morphemes (*i.e.*, a Morpheme-Based annotation strategy) or whether we should preserve their special orthographic status within the trees (*i.e.*, a Word-Based annotation strategy). In this paper we empirically address this challenge in the context of the development of Language Resources for Modern Hebrew. We compare and contrast the Morpheme-Based and Word-Based annotation strategies of pronominal clitics in Modern Hebrew and we show that the Word-Based strategy is more adequate for the purpose of training statistical parsers as it provides a better PP-attachment disambiguation capacity and a better alignment with initial surface forms. Our findings in turn raise new questions concerning the interaction of morphological and syntactic processing of which investigation is facilitated by the parallel treebank we made available.

1. Introduction

The development of statistical parsing models for different languages utilizing an annotated corpus for training syntactic analyzers/disambiguators has become increasingly popular during the last decade following the success of statistical parsers developed for English (Collins, 2003; Charniak, 1997; Bod et al., 2003; Charniak and Johnson, 2005), trained and tested on the Wall Street Journal (WSJ) standard benchmark corpora (Marcus et al., 1994). Such efforts typically involve the development of a body of annotated text (a 'treebank', where tree structures represent syntactic structures of phrases and sentences), followed by an application of a parsing model to the resulting treebank (*e.g.*, (Bikel and Chiang, 2000; Dubey and Keller, 2003)).

The annotation of newly developed corpora is typically inspired by the annotation scheme of the WSJ Penn Treebank (Marcus et al., 1994) yet annotators of texts in a different language often face the need to deviate from the guidelines for annotating English. Even for one and the same language it was shown that representational variations significantly affect parsing accuracy (Johnson, 1998; Klein and Manning, 2003), let alone varying the representation between parse-trees in different languages. So, the question of what information to encode is always accompanied with the question of how to represent such information in order to optimize performance on the task we have in mind.

Morphologically rich languages pose a challenge to the annotators of syntactic treebanks in terms of the status of orthographic (space-delimited) words in the syntactic parse trees (Sima'an et al., 2001; Maamouri et al., 2004). In such languages a single word may carry different sorts of information (Adler and Elhadad, 2006; Bar-Haim et al., 2005) and the different morphs composing a word may stand for, or indicate a relation to, other elements in the syntactic parse tree (Tsarfaty, 2006). When annotating syntactic tree structures the question arises whether we should represent a word as a sequence of morphs belonging to distinct morphosyntactic categories or whether we should preserve the special status of orthographic (space-delimited) words while providing the additional morphological information by other means.

The status of words in morphologically rich languages has already been subject to theoretical debates between linguists working in different morphological schools. Post Bloomfieldian *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). In *Word-Based (WB)* approaches (Blevins, 2006) words are considered the atomic units of the language, and morphological considerations reflect generalizations about their syntactic behavior.[1] The WB vs. MB debate also begs a question concerning the relation between syntax and orthography — to what extent do orthographic units reflect syntactic structures? Do orthographic units correspond to the yield of the syntactic tree (WB) or are they better split-off into several separate leaves (MB)?

In this paper we address the empirical consequences of this theoretical challenge in the context of the development of Language Resources for Semitic Languages. Specifically, we discuss and empirically demonstrate the adequacy of a Word-Based (WB) annotation strategy for pronominal suffixes in Modern Hebrew (henceforth Hebrew). Pronominal suffixes in Hebrew may attach to function words such as prepositions and case markers to indicate their pronominal complements via a set of inflectional features (such as gender, number and person). Here we compare and contrast MB and WB annotation strategies of such forms and empirically evaluate them on parallel versions we developed of the Hebrew Treebank.

Our quantitative and qualitative analysis shows that the WB

[1] In Psycholinguistics, debates about the structure of the mental lexicon show similar concerns (Ravid, 2006).

strategy is more adequate than the MB strategy for statistical parsing as it provides better PP attachment disambiguation capacity of the resulting treebank grammar and is more faithful to the surface forms we begin with. Our findings in turn raise new questions (Section 4) concerning the interaction between morphological and syntactic processing of which investigation is facilitated by the new parallel corpus we provide.

2. The Data

Words in languages of the Semitic family, such as Hebrew and Modern Standard Arabic (Arabic), have a rich morphological structure. A single orthographic space-delimited word (henceforth, a 'word') in Hebrew may contain different sorts of information including the root/template deriving the stem, agreement features such as tense, number and gender marked by inflectional morphology, and additional prefixes marking prepositions, relativizers, and conjunction concatenated onto the stem. Such multifaceted analysis of a word is illustrated in (1).[2]

(1) *w-kf-n-rdm-ti*
and-when-middle-sleep-1pers.sing
'and when I fell asleep'

The general strategy currently employed by Hebrew NLP resource developers is to identify a single stem within the word (typically, from an open class category), and then distinguish the morphological material internal to the stem from the morphological material external to it. The 'internal' material is encoded on top of the respective syntactic category,[3] and the external morphs are segmented away and get assigned their own Part of Speech (POS) tags. Such a strategy is syntactically justified since elements such as prepositions, relativizers and conjunction markers typically attach higher and are dominated by a different parent than the one dominating the stem (Tsarfaty, 2006).

The main source of debate between different developers of Hebrew language resources has to do then with the distinction between morphological material internal to the stem and morphological material external to it. A canonical case of disagreement is the case of Hebrew pronominal suffixes. Pronominal suffixes in Hebrew may attach to function words such as prepositions, accusative markers and possessive markers to mark their pronominal complements, as illustrated in (2).[4] The analysis of pronominal suffixes in the Hebrew Treebank is Morpheme-Based whereby such forms are segmented into two distinct elements, one a generic preposition, and the other a full-fledged pronoun carrying its own inflectional features. Each of these elements is then represented as a distinct leaf in the syntactic tree. The resulting respective Treebank yields are thus illustrated in (3). (It is to note that only the yields in (2) correspond to the actual surface realization of such forms in Hebrew, whereas the yields in 3 impose additional morphological decomposition.)

(2) a. **Prepositions:**
hwa ba **al***i*
he came **to**.1p.sing
He came to me

b. **Possessive Marker:**
hildim **fl***nw*
the-children **of**.1p.plural
Our children

c. **Accusative Marker:**
hwa rah **awt***nw*
he saw ACC.1p.plural
He saw her

(3) a. **Prepositions:**
hwa ba **al** *ani*
he came **to** I
He came to me

b. **Possessive Marker:**
hildim **fl** *anxnu*
the-children **of** we
Our children

c. **Accusative Marker:**
hwa rah **at** *anxnw*
he saw ACC we
He saw us

Prepositions/Markers segmented away from cliticized elements clearly share properties with the respective bare prepositions/markers, namely that they all require a Noun Phrase to form a saturated prepositional (or otherwise marked) phrase. Yet we suggest that the former exhibit a slightly different behavior. While bare prepositions may attach to complex Noun Phrases (such as modified Nouns or Construct-State Nouns), the segmented prepositions subcategorize only for pronouns. Upon selecting a "light" Pronoun these prepositions form phrases that manifest syntactic behavior distinct from that of prepositional phrases saturated with other types of NPs. For instance, they cannot undergo the dative shift, as illustrated in the minimal pair (4)–(5).

(4) a. *ntti* **l***w* *mtnh*
gave.1p.sing **to**.3p.masc.sing a-present
I gave him a present

b. **ntti* *mtnh* **l***w*
*gave.1p.sing a-present **to**.3p.masc.sing
*I gave a present to him

(5) a. *ntti* **l***ild* *mtnh*
gave.1p.sing **to**-the-child a-present
I gave the child a present

b. *ntti* *mtnh* **l***ild*
gave.1p.sing a-present **to**-the-child
I gave a present to the child

Further, Prepositions segmented away from cliticized elements cannot scope over a conjoined Noun Phrase, while bare prepositions can. Thus, "light" (cliticized) phrases may only conjoin with fully saturated prepositional phrases, as illustrated in (6)-(7) for the preposition **l** (for). This behavior can in turn be replicated for any of the above mentioned phrase markers.

[2] We use the transliteration scheme proposed in (Sima'an et al., 2001) throughout.

[3] Broadly understood as a combination of syntactically and morphologically relevant information, *e.g.*, a combination Part-of-Speech tags and agreement features as used in (Adler and Elhadad, 2006).

[4] We consider such affixes simple clitics in the sense of (Zwicky, 1977, page 10, 4.1(b)).

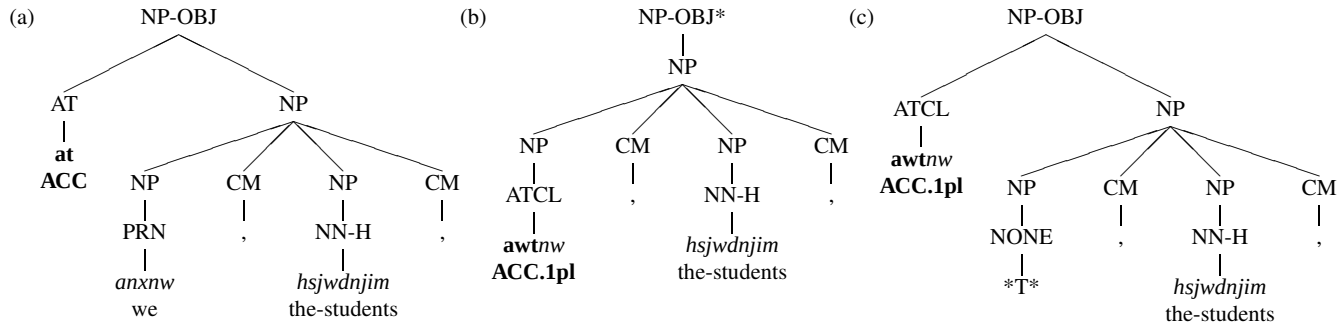

Figure 1: Cliticized Elements in Apposition Structures: (a) treats apposition in the Morph-based strategy, (b) shows an erroneous Word-Based analysis, and (c) illustrates our proposed remedy for the Word-Based treatment using traces.

(6) a. **l**i*wab* *wrewt*
for-Yoav and-Reut
for Yoav and Reut

b. ***l**i* *wrewt*
***for**.1p.sing and-Reut
*for me and Reut

(7) a. **l**i* *wliwab*
for.1p.sing and-**for**-Yoav
for me and for Yoav

b. **l**i* *w**l**w*
for.1p.sing and-**for**.3p.sing.masc
for me and for him

3. The Proposal

The Hebrew Treebank was originally annotated according to an MB strategy. The main motivation for adopting an MB annotation strategy for pronominal clitics in Hebrew is the discrepancy found between constituent boundaries and the boundaries of such cliticized words. Let us take, for instance, the Noun Phrase in (8).

(8) *awtnw, hsjwdnjim,*
we.ACC, the-students,
us, the students,

In Figure (1a) the accusative marker is a sibling of the elaborated apposition structure of the NP and licenses it as a direct object. Taking the accusative marker and the pronoun "we" as a word that occupies a single subconstituent (Figure (1b)) would deem the structure ungrammatical.

On the other hand, the MB annotation strategy poses a problem for any automatic parsing system: an additional non-trivial morphological disambiguation stage needs to take place prior to parsing.

We claim that the aforementioned discrepancies need not impose a MB strategy (with its implied additional computational complexity), and may be accounted for within WB annotation strategies as well. In what follows we put forth one concrete proposal to do so, and show that our WB strategy is not only theoretically adequate, but also empirically superior to the original MB one.

Taking the Morpheme-Based (MB) Hebrew Treebank analysis as a baseline, we propose an alternative Word-Based (WB) analysis of pronominal clitics in the Treebank as inflectional features on top of specialized categories of "cliticized prepositions/possesives/case markers". The specialized tags capture membership in a class of prepositions/markers that share a distinct syntactic behavior, and the features are understood as indicating agreement with a pronoun which can be dropped on pragmatic grounds. This analysis is in line with treating pro-dropped elements across languages and we similarly indicate them as traces marking empty elements. Thus, Figure (1c) shows how our trace analysis remedies the ungrammaticality of a naïve WB annotation proposal.

We illustrate the resulting competing analyses on our sample sentences in Figure (2), where (a) corresponds to the current MB analyses, and (b) provides our alternative WB treatment. We further note that the yield of the WB analyses in (b) always corresponds directly to the surface sequence, while the MB analyses in (a) presuppose a preceding morphological analysis and segmentation stage.[5] We consider the direct correspondence to the input an advantage for the WB strategy when parsing moprhologically rich languages especially in the context of joint morphological-syntactic disambiguation frameworks as argued for in (Tsarfaty, 2006; Cohen and Smith, 2007; Goldberg and Tsarfaty, 2008).

4. Experimental Setup

Goal We set up a series of experiments to compare and contrast the adequacy of the MB and WB annotation strategies for Hebrew pronominal clitics by evaluating the parsing performance of two different PCFG-based treebank grammars trained on parallel treebanks of which the trees correspond to either of the annotation strategies.

Data The data is taken from the Modern Hebrew Treebank version 1.0 (Sima'an et al., 2001), which consists of 5000 sentences from the daily newspaper 'Ha'aretz' annotated with integrated morphological and syntactic representations. We split the data into a Test-Set and a Train-Set, where the Test-Set constitutes the first 500 non-empty sentences.

Tag-Set For the purpose of our experiments we extracted bare tree-skeletons in which syntactic categories are extended with a handful of morphosyntactic features that have proven successful in increasing the disambiguation capacity in parsing Hebrew (Tsarfaty, 2006; Tsarfaty and Sima'an, 2007). Specifically, we strip-off inflectional

[5] Note that this stage is non-deterministic and non-trivial, as illustrated by, *e.g.*, (Adler and Elhadad, 2006; Bar-Haim et al., 2005).

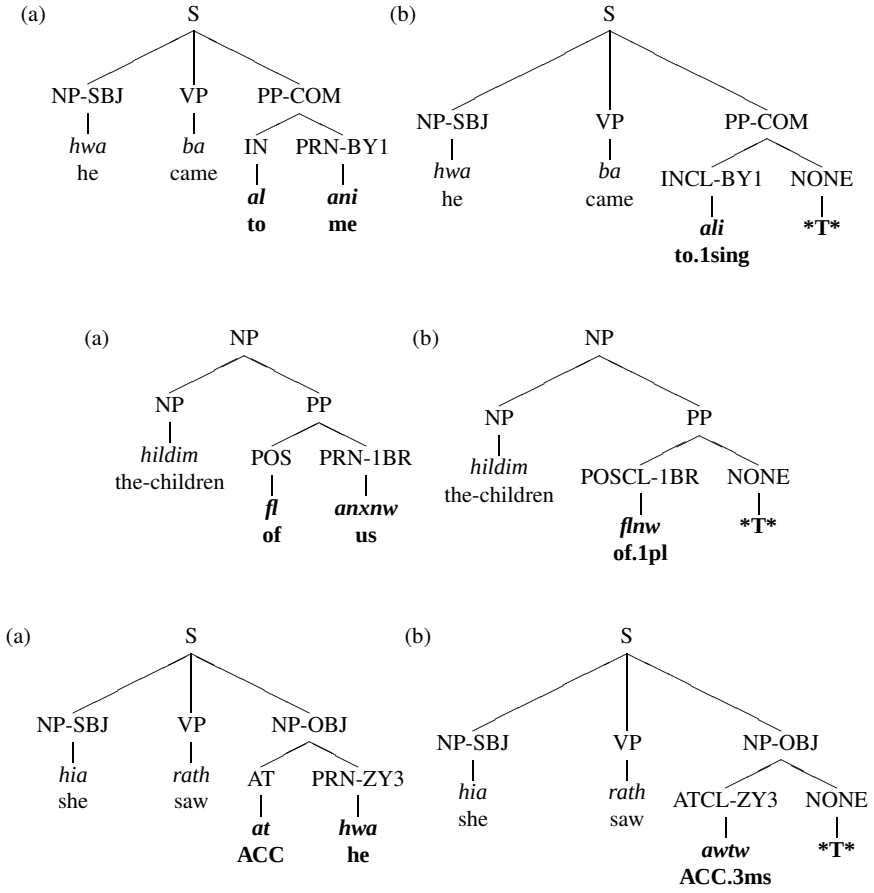

Figure 2: Morpheme-Based and Word-Based Annotation Strategies: (a) trees illustrate the Morph-Based strategy as used in the Hebrew TB v1.0, and corresponding (b) trees illustrate our proposed Word-Based analysis.

features (gender, number) and functional features (SUBJ, OBJ), but retain the distinction between definite and indefinite Nouns (NNH/NN) and Noun-Phrases (NPH/NP), and between finite and non-finite Verbs (VB/VBM) and Verb-Phrases (VP/VPINF). Following (Goldberg and Elhadad, 2007) we also distinguish Possessive Prepositional Phrases (PPPs) from ordinary Prepositional Phrases (PPs).

Procedure We implemented software that converts the default Morpheme-Based Treebank analyses in Figure (2a) to the Word-Based analyses in Figure (2b). We collapse cliticized pronouns onto prepositions, accusative markers, and possessive markers, and for each conversion we train a PCFG on instances of the Treebank before and after the conversion. We then use an efficient general-purpose parser, Bitpar (Schmid, 2004), to parse unseen sentences with the resulting Treebank grammars and strip off our morphological features for the purpose of evaluation. In order to isolate structural representation effects on the disambiguation capacity of the treebank grammars from morphological disambiguation matters we make sure that the test sentences are morphologically segmented (when applicable) and tagged correctly prior to parsing.

Evaluation Comparing the performance of the parser for different annotation strategies is not a trivial matter as the WB annotation results in sentences that are shorter in terms of POS sequences. (This can be seen by comparing, *e.g.*, the difference in length of the strings in our examples in (2) and their corresponding (MB) treebank yields in (3).) Thus, a plain comparison of corpus-averaged PARSEVAL measures would not be informative. In order to compare the performance of the parser on different annotation strategies we first quantitatively compare the PARSEVAL measures averaged on sentences that have not been changed by the conversion, which would give us an indication of the disambiguation capacities of the treebank grammar for bare prepositions and pronouns. We then qualitatively analyze differences in the resulting parse-trees for either strategy and contrast their (dis)advantages.

5. Results and Analysis

Table 2 shows the Labeled Precision, Labeled Recall and F-Measure results of parsing all trees that were not affected by the conversion (*No Clitics*) and of a subset of all the trees that were not affected by the conversion yet contained a bare preposition (*Bare Prepositions*). The results of parsing with PCFGs obtained before and after the conversion show the same or slightly decreased performance (in a small rate) for the WB strategy. This means that our conversion doesn't have a significant influence on the disambiguation capacity

	All Sentences	Train-Set	Test-Set
Inflected Prepositions	891	801	90
Inflected Possessive Markers	188	165	23
Inflected Accusative Markers	169	151	18

Table 1: Corpus Statistics: Pronominal Clitics in the Hebrew Treebank v1

	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)

Table 2: Parsing Results on (subsets of the) Test-Set: Averaged Labeled Precisions/Labeled Recall and (F-Measure)

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

Table 3: Comparison of the resulting parses for the WB and MB analyses on cliticized prepositions in the test set

for structures that do not involve cliticized elements, which we attribute to the high frequency of bare prepositions and independent pronouns in the Treebank.

For the sentences in which cliticized elements were converted from the MB analysis to the WB analysis, a manual comparison of the resulting parse trees (Table 3) reveals that from the 131 occurrences of cliticized prepositions in the test set, 116 received an identical analysis from either annotation strategy. Of the 15 differing analyses, 5 occurrences are parsed correctly under the WB scheme but not under the MB scheme, 8 are assigned an incorrect structure under both schemes but the structure under the WB scheme is more acceptable (higher overlap with the gold tree), and the remaining 2 are assigned an equally unacceptable analyses under both schemes. That is, on the sentences affected by the conversion, the WB analysis is almost always as good as, and often better than, the original MB analysis. Figure 3 illustrates a tree fragment that was disambiguated correctly under the WB representation, but not under the MB representation.

Our qualitative analysis shows that the main source of errors for the MB strategy is its tendency to learn high attachment for prepositions that originate from cliticized elements. Under the MB analysis these prepositions share a probability distribution with bare prepositions and therefore tend to attach high to NPs with elaborated internal structures. The WB analysis constrains such prepositions to select a single pronoun only and form a "light" prepositional phrase. This provides better alignment with the gold constituent structure, with better chances of separating out subsequent constituents accordingly.

This advantage is a consequence of the fact that the WB strategy relieves the parser from the duty to disambiguate an attachment to independent elements that are not there in the surface form to begin with. For cases in which the tree structure requires interaction of the cliticized pronoun with coordinated Noun-Phrases (using traces), our qualitative analysis shows that neither of the strategies recovered the correct analysis. We conjecture that methods for recovering traces such as (Levy and Manning, 2004; Schmid, 2006) would be more appropriate for the treatment of such structures, more so than imposing a shared distribution on morphologically distinct elements.

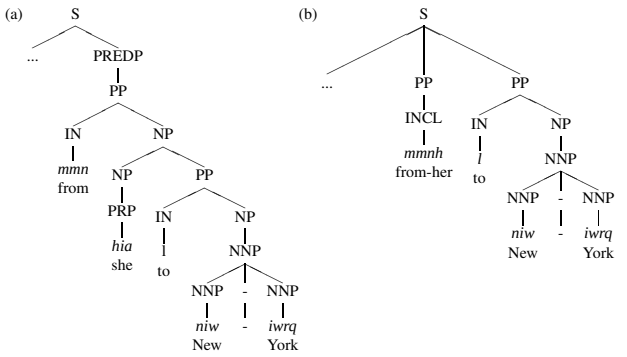

Figure 3: A tree fragment in which the parser failed to recover the correct attachment under the MB grammar (a), but did recover it correctly under the WB grammar (b). In (a) the preposition *mmn* (from), originating from the cliticized element *mmnh* (from-her) is attached to internally complex NP, a structure that is not licensed by the Modern Hebrew Grammar. In (b) the cliticized element is explicit in the annotation and the PP is attached accordingly.

6. Discussion and Conclusion

Annotating syntactic structures for morphologically rich languages requires annotators to make decisions concerning the status of morphologically complex surface forms in the syntactic parse trees. Specifically, in Hebrew there is a discrepancy between possible syntactic analyses for pronominal clitics imposed by Morpheme-Based and Word-Based Annotation Strategies. Through a quantitative and qualitative analysis of the parsing results under the two competing annotation strategies we have shown that the Word-Based strategy is advantageous to the Morpheme-Based one as it provides better disambiguation of PP attachment. The WB strategy is further more faithful to the surface form as it maintains the status of orthographic units in the yield of the syntactic parse-tree. So far we have experimented with sentences for which correct POS tagging was assumed prior to parsing and we conjecture that the WB strategy will have further advantages in a more realistic scenario in which a stage of word segmentation and POS tagging is assumed to precede (or take place jointly with) the parsing process. In the case of pronominal clitics, prior morphological decomposition will simply be unnecessary.

In addition to the empirical analysis, our work results in the availability of parallel corpora for Hebrew in which pronominal clitics are annotated according to either Morph-Based or Word-Based strategy, which in turn facilitates the empirical exploration of emerging follow up questions. For instance, it would be interesting to check whether the advantage of Word-Based strategies persists with more sophisticated (*e.g.*, lexicalized) parsing models. We further suggest that such annotation decisions may and should be empirically evaluated in the context of other languages as well (*e.g.*, Arabic)[6] yet we leave the investigation of the cross-linguistic angle of rich morphosyntactic representations for future research. Finally, the results of these and similar investigations will facilitate fine-tuning of the division of labor between a morphological and a syntactic components in joint disambiguation frameworks as proposed in (Tsarfaty, 2006; Cohen and Smith, 2007; Goldberg and Tsarfaty, 2008).

Acknowledgments

We thank Remko Scha, Khalil Sima'an and Jelle Zuidema from the Univrsity of Amsterdam (Netherlands) and Meni Adler and Michael Elhadad from Ben-Gurion University (Israel) for comments and discussion. We further gratefully acknowledge James P. Blevins and his Morphology course given at the LSA institute 2007 (Stanford University) for introducing us with the Word-Based linguistic perspective on morphology. The work of the first author in Amsterdam as well as collaboration visits to Israel were financed by NWO, grant number 017.001.271. The work of the second author is partially supported by the Lynn and William Frankel Center for Computer Sciences.

[6] We suggest that the methodology and analysis we use here are immediately applicable to other Semitic Languages, yet we cannot properly illustrate it under the page limit.

7. References

M. Adler and M. Elhadad. 2006. An Unsupervised Morpheme-Based HMM for Hebrew Morphological Disambiguation. In *Proceedings of COLING-ACL 2006*.

R. Bar-Haim, K. Sima'an, and Y. Winter. 2005. Choosing an Optimal Architecture for Segmentation and POS-Tagging of Modern Hebrew. In *ACL Workshop on Computational Approaches to Semitic Languages*.

D. Bikel and D. Chiang. 2000. Two Statistical Parsing Models Applied to the Chinese Treebank. In *Second Chinese Language Processing Workshop*, Hong Kong.

J. P. Blevins. 2006. Word-Based Morphology. *Journal of Linguistics*, 3(42):531–573.

L. Bloomfield. 1933. *Language*. University of Chicago Press.

Rens Bod, Remko Scha, and Khalil Sima'an, editors. 2003. *Data-Oriented Parsing*. CSLI Publications.

Eugine Charniak and Mark Johnson. 2005. Coarse-to-fine N-Best Parsing and Maxent Discriminative Reranking. In *Proceedings of ACL*.

E. Charniak. 1997. Statistical Parsing with a Context-Free Grammar and Word Statistics. In *AAAI/IAAI*, pages 598–603.

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

M. Collins. 2003. Head-Driven Statistical Models for Natural Language Parsing. *Computational Linguistics*.

A. Dubey and F. Keller. 2003. Probabilistic Parsing for German using Sister-Head Dependencies. In *Proceedings of ACL*.

Y. Goldberg and M. Elhadad. 2007. Toward Better Understanding of Hebrew NP Chunks. In *ISCOL/BIFSAI 2007*.

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).

M. Johnson. 1998. PCFG Models of Linguistic Tree Representations. *Computational Linguistics*, 24(4):613–632.

D. Klein and C. Manning. 2003. Accurate Unlexicalized Parsing. In *Proceedings of ACL*, pages 423–430.

R. Levy and C. Manning. 2004. Deep Dependencies from Context-Free Statistical Parsers: Correcting the Surface Dependency Approximation. In *Proceedings of ACL*.

M. Maamouri, A. Bies, T. Buckwalter, and W. Mekki. 2004. The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus.

M. Marcus, G. Kim, M. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz, and B. Schasberger. 1994. The Penn Treebank: Annotating Predicate-Argument Structure.

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

D. Ravid. 2006. Word-level Morphology: A Psycholinguistic Perspective on Linear Formation in Hebrew Nominals. *Morphology*, 16(1):127–148.

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

H. Schmid. 2006. Trace Prediction and Recovery with Unlexicalized PCFGs and Slash features. 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 and K. Sima'an. 2007. Three-Dimensional Parameterization for Parsing Morphologically Rich Languages. In *Proceedings of IWPT*.

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

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