The Interplay of Syntax and Morphology in Building Parsing Models for Modern Hebrew

Reut Tsarfaty
Institute for Logic, Language and Computation, University of Amsterdam
rtsarfat@science.uva.nl

Abstract. As of yet, there is no statistical parser for Modern Hebrew (MH). Current practice in building parsing models is not immediately applicable to languages that exhibit strong interaction between syntax and morphology, e.g. Modern Hebrew, Arabic and other Semitic languages. We suggest that incorporating morphological and morphosyntactic information into the parsing model is essential for parsing Semitic languages. Using a morphological analyzer, a part-of-speech tagger, and a PCFG-based general purpose parser, we segment and parse unseen MH sentences using a small annotated corpus. The Parseval scores obtained are not comparable to those of, e.g., state-of-the-art models for English, due to remaining syntactic ambiguity and limited morphological treatment. We conjecture that adequate morphological and syntactic processing of MH should be done in a unified framework in which morphology and syntax can freely interact and share information in both directions.

1 Introduction

The structure of Semitic languages poses clear challenges to the traditional view of Natural Language Processing, in which different processing layers[1] are handled separately. Specifically, Semitic languages demonstrate strong interaction between morphological and syntactic processing, which limits or precludes the application of standard tools and techniques for parsing Semitic languages.

The problem, in essence, is as follows. Modern Hebrew (MH), Arabic, and other Semitic languages, have a rich morphology. Affixes that are appended to the stem of a word carry substantial information and serve different syntactic functions. Therefore, a first step towards utterance understanding is to extract the different constituents that exist at the word level to allow for further processing (e.g., parsing). However, because of the large-scale morphological ambiguity in Semitic languages already at the word level, and due to the lack of vocalization in written texts, each word-form may have multiple possible morphological analyses. Picking out the correct analysis is largely dependent on contextual information, which may be carried over syntactic structures. Therefore, a suitable treatment of morphological analysis in Semitic languages demands a treatment of syntactic analysis and vice versa.

[1] I.e., phonological, morphological, syntactic, semantic and pragmatic.

Proceedings of the Eleventh ESSLLI Student Session
Janneke Huitink & Sophia Katrenko (editors)

This work focuses on MH and presents a baseline architecture for parsing that incorporates one level of morphological processing, namely morphological segmentation. The particular contribution of this work is to demonstrate that MH statistical parsing is feasible, even with a relatively small set of annotated data. Yet, in the current setting, our results fall behind those achieved for, e.g., English, which may be due to corpus size, annotation scheme, limited morphological treatment, and flexible sentence structure. In the future we intend to develop models that implement a closer interaction between morphological and syntactic processing, which are better suited for capturing linguistic phenomena in Semitic languages, and are expected to boost MH parsing accuracy.

2 Linguistic Data

2.1 Semitic Morphology

Morphological analysis of a MH word consists of, at least, the stem, prefixes, person, number and gender inflections, pronominal suffixes, and so on (Segal(2000); Bar-Haim(2005); Sima'an et al.(2001)Sima'an, Itai, Winter, Altman, and Nativ). The different morphological processes that take place in the formation of MH words can be roughly divided into (i) derivational morphology, (ii) inflectional morphology and (iii) concatenation.

Verbs, nouns, and adjectives in Semitic languages are derived from (tri-)consonantal roots plugged into templates of consonant/vowel skeletons. The lexical items in (1), for example, are all derived from the same root, $[i][l][d]$.[2] ('...' indicates surface forms, $[c]$ indicates template's slots for root's consonants, (c) indicates doubling of root's consonants.)

	a.	'ild'	b.	'iild'	c.	'mwld'
(1)		$[i]e[l]e[d]$		$[i]i[l](l)e[d]$		$mu[][l](l)a[d]$
		a child (n)		deliver a child (v)		innate (adj)

In addition, MH has a rich array of agreement features expressed at the word level. Features such as gender, number and person are expressed in the word's inflectional morphology. Verbs, adjectives, determiners and even numerals have to agree on the inflectional features with the noun they complement or modify. For example, in (2b) the suffix *heh* (h) alters the noun 'ild' (child) and its modifier 'gdwl' (big) to feminine gender.

	a.	ild gdwl	b.	ildh gdwlh
(2)		child.MS big.MS		child.FS big.FS
		a big boy		a big girl

Finally, many particles in MH, such as conjunctions, prepositions, complementizers and relativizers, are prefixed to the word. Such particles serve syntactic functions that are distinct from that of the stem, yet a multiplicity of them may be concatenated together with the stem to form a single (space-delimited) word. For example, the word form in (3) is formed from a conjunction, a relativizer, a preposition, and a definite noun phrase.

[2]The transliteration we use is adopted from (Sima'an et al.(2001)Sima'an, Itai, Winter, Altman, and Nativ) and repeated in the appendix for convenience.

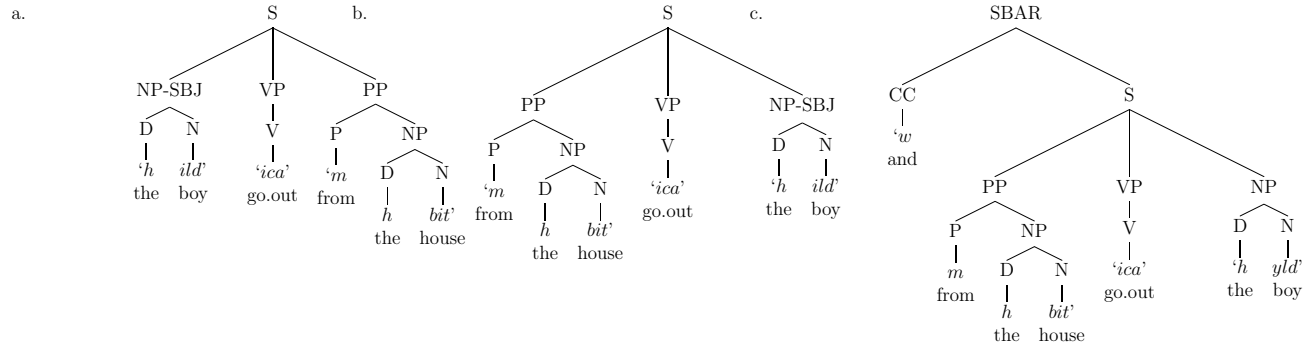

Figure 1.1: Syntactic Structures of MH Phrases (‘. . . ’ mark word boundaries)

(3) a. ‘wksmhbit’
w ks m h bit
and when from the house

Identifying such constituents within words is crucial for analyzing the syntactic structure of sentences, as they reveal structural dependencies such as subordinate clauses, adjuncts, and prepositional phrase attachment.

2.2 Syntactic Structures

Turning now to syntactic structures in MH, we first note that sentences in MH have a relatively free word order.[3] In general, MH allows for both SV and VS, and in some circumstances for SVO permutations such as VSO and others (Shlonsky(1997)). To illustrate, figures 1.1a–1.1b show two distinct syntactic structures that express the same grammatical relations.

Further, as a result of the concatenation process the constituents that are combined to form phrases and sentences in MH are not words, but rather, the morphological constituents that were concatenated together to form words. Figure 1.1c demonstrates that a MH word-form may coincide with a single constituent, as in ‘ica’ (leave, go out), it may overlap with an entire phrase, as in ‘h ild’ (the boy), or it may span across phrases as in ‘w m h bit’ (and from the house). Thus, it becomes clear that in order to perform syntactic analysis (parsing) of MH sentences we must first set the sequence of morphological constituents in place.

2.3 The Problem: Ambiguity

MH and other Semitic languages exhibit a large-scale ambiguity at the word level. This means that there are multiple ways in which a word can be broken down into its constituent morphemes. This is further complicated by the fact that most vocalization marks (diacritics) are omitted in MH texts. The word-form ‘fmnh’, for instance, has four readings, to

[3]Relative to, e.g., English.

Segmentation:	fmnh	fmnh	fmnh	fmnh	f + mnh
Vocalization:	shmena	shamna	shimna	shimna	she + mana
Analysis:	fat.FS	grew-fat.FS	lubricate.FS	oil-of.FS	that + counted
Meaning:	fat (adj)	grew fat (v)	lubricate (v)	her oil (n)	that counted (rel)

Table 1.1: Morphological Analyses of the Word Form 'fmnh'

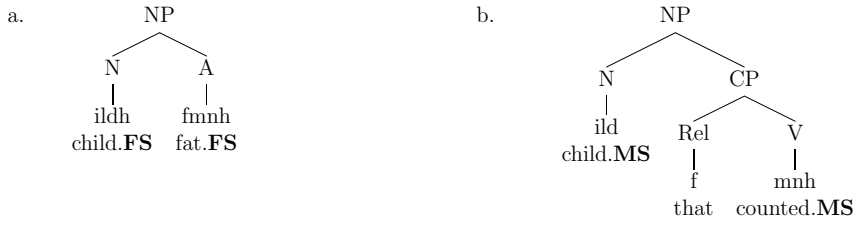

Figure 1.2: Morphological Ambiguity Resolution in Different Syntactic Contexts

which (at least) five morphological analyses can be found, as shown in table 1.1.[4] Moreover, the different morphological analyses of a word may give rise to different segmentation possibilities. In the case of the word-form 'fmnh' the five morphological analyses correspond to two distinct morphological segmentation possibilities, as observed in the table.

The morphological analysis of a word-form, and in particular its morphological segmentation, cannot be disambiguated without reference to context, i.e., an utterance. When context is available, various syntactic features of surrounding forms provide useful hints for choosing the correct analysis. Figures 1.2a–1.2b show the correct analyses of the form 'fmnh' in the different syntactic contexts in which they appear. Note that the correct morphological analysis maintains agreement on gender (M/F) and number (S/P) between the noun and the verb or the adjective. In particular, the analysis 'that counted' is easily picked out for 1.2b as it is the only one maintaining agreement with the modified noun.

Therefore, we would want to conclude that syntactic processing (parsing) must precede morphological analysis; however, this would be in apparent contradiction to our previous conclusion. For this reason, independent morphological and syntactic mechanisms for MH will not suffice. In what follows we describe a parsing architecture that incorporates one level of morphological processing, namely segmentation, as a first attempt to model the interaction between morphological and syntactic processing. We further observe that the morphosyntactic categories that are assigned to morphological segments must coincide with the lowest level of non-terminals in the syntactic parse tree. Therefore, we incorporate an intermediate level of processing, part-of-speech (POS) tagging, to ensure *agreement* between the morphological and the syntactic tasks.

[4]In fact, a statistical study on a MH corpus has shown that the average number of possible analyses per word-form was 2.1, while 55% of the word-forms were morphologically ambiguous (Sima'an et al.(2001)Sima'an, Itai, Winter, Altman, and Nativ).

3 Formal Settings

Before describing our baseline architecture, we first develop a formal account of an integrated model for morphological and syntactic processing in a generative probabilistic framework.

Let w_1^m be a sequence of words from a fixed vocabulary (i.e., a sequence of surface word-forms as they occur in the text), let s_1^n be a sequence of segments of words from a (different) vocabulary, let t_1^n be a sequence of morphosyntactic categories from a finite tag set, and let π be a syntactic parse tree.

We define morphological *segmentation* as the task of identifying the sequence of morphological constituents that were concatenated to form a sequence of words. Formally, we define the task as (1.1), where $seg(w_1^m)$ is the set of segmentations resulting from all possible morphological analyses of the words.

$$s_1^{n*} = \underset{s_1^n \in seg(w_1^n)}{arg\max}\, P(s_1^n|w_1^m) \tag{1.1}$$

Syntactic analysis, *parsing*, is the task of identifying the structures of phrases and sentences. In MH, such tree structures combine segments of words that serve different syntactic functions. Formally, we define it as (1.2), where $yield(\pi)$ is the ordered set of leaves of the syntactic parse tree.

$$\pi^* = \underset{\pi \in \{\pi' : yield(\pi') = s_1^n\}}{arg\max}\, P(\pi|s_1^n) \tag{1.2}$$

The *part-of-speech (POS) tagging* task is concerned with assigning morphosyntactic categories to words. Following our theoretical exposition in section 2, it becomes clear that in MH categories are assigned to morphological segments rather than to words. So we define the task of POS tagging as (1.3), where $analyses(s_1^n)$ is the set of possible POS tags' assignments for a sequence of morphological segments.

$$t_1^{n*} = \underset{t_1^n \in analyses(s_1^n)}{arg\max}\, P(t_1^n|s_1^n) \tag{1.3}$$

The task of the *integrated model* for morphological and syntactic processing is to find the most probable morphological segmentation *and* syntactic parse tree given a sequence of word-forms, as in (1.4).

$$\langle \pi, s_1^n \rangle^* = \underset{\langle \pi, s_1^n \rangle}{arg\max} P(\pi, s_1^n|w_1^m) \tag{1.4}$$

We can rewrite (1.4) using conditional probabilities, thus distinguishing the morphological and syntactic tasks, yet conditioning the latter on the former.

$$\langle \pi, s_1^n \rangle^* = \underset{\langle \pi, s_1^n \rangle}{arg\max} \underbrace{P(\pi|s_1^n, w_1^m)}_{parsing} \underbrace{P(s_1^n|w_1^m)}_{segmentation} \tag{1.5}$$

In order to ensure agreement between the morphological and syntactic tasks, we incorporate an intermediate level of POS tagging into the model, which ensures that the

morphosyntactic categories assigned to the morphological segments coincide with the lowest level of non-terminals in the syntactic parse trees (cf. (Charniak et al.(1996)Charniak, Carroll, Adcock, Cassandra, Gotoh, Katz, Littman, and McCann)). This results in (1.7).

$$\langle \pi, t_1^n, s_1^n \rangle^* = \underset{\langle \pi, t_1^n, s_1^n \rangle}{arg\max} P(\pi, t_1^n, s_1^n | w_1^m) \tag{1.6}$$

$$= \underset{\langle \pi, t_1^n, s_1^n \rangle}{arg\max} \underbrace{P(\pi | t_1^n, s_1^n, w_1^n)}_{parsing} \underbrace{P(t_1^n | s_1^n, w_1^n)}_{tagging} \underbrace{P(s_1^n | w_1^m)}_{segmentation} \tag{1.7}$$

Finally, we employ the assumption that $P(w_1^m | s_1^n) \approx 1$, since morphological segments can only be conjoined in a certain order.[5] So, instead of (1.5) and (1.7) we end up with (1.8), (1.9) respectively.

$$\approx \underset{\langle \pi, s_1^n \rangle}{arg\max} \underbrace{P(\pi | s_1^n)}_{parsing} \underbrace{P(s_1^n | w_1^m)}_{segmentation} \tag{1.8}$$

$$\approx \underset{\langle \pi, t_1^n, s_1^n \rangle}{arg\max} \underbrace{P(\pi | t_1^n, s_1^n)}_{parsing} \underbrace{P(t_1^n | s_1^n)}_{tagging} \underbrace{P(s_1^n | w_1^m)}_{segmentation} \tag{1.9}$$

4 Evaluation Metrics

The intertwined nature of morphology and syntax in MH also challenges standard parsing *evaluation metrics*, as the proposed segmentation need not coincide with the gold segmentation for a given sentence. Therefore, we cannot use morphemes as the basic units for comparison. Since words are complex entities that can span across phrases, we cannot use them for comparison either. Therefore, we redefine *precision* and *Recall* by considering the spans of syntactic categories based on the (space-free) sequences of characters they correspond to. Formally, we define syntactic constituents as $\langle i, A, j \rangle$ where i, j mark the location of characters, we define $T = \{\langle i, A, j \rangle | A \textit{ spans from } i \textit{ to } j\}$ and $G = \{\langle i, A, j \rangle | A \textit{ spans from } i \textit{ to } j\}$ as the test/gold parse trees respectively, and calculate as follows.

$$\textit{labeled precision} = \frac{\#(G \cap T)}{\#T} \tag{1.10}$$

$$\textit{labeled recall} = \frac{\#(G \cap T)}{\#G} \tag{1.11}$$

[5] In MH, conjunctions, relativisers, prepositions and definite markers must be attached in front of the stem, pronominal and inflectional affixes appear at the end of the stem, and derivational morphology shows up inside the stem. Thus, a sequence of morphological segments can only be conjoined in a certain order. To illustrate, although the MH form 'hkph' is ambiguous between three morphological analyses; (i) 'h'+'kph' (the + coffee) (ii) 'hkph' (lap, surrounding) and (iii) 'hkp'+'h' (perimeter + of-her), restoring the surface forms that correspond to the different sequences in (i)–(iii) must result in the word-form 'hkph'.

5 Experimental Setup

5.1 The Baseline Architecture

Our departure point for the syntactic analysis of MH is that the basic units for processing are not words but the morphological segments that are concatenated together to form words. Therefore, we obtain a segment-based probabilistic grammar by training a probabilistic context-free grammar on a segmented and annotated MH corpus (Sima'an et al.(2001)Sima'an, Itai, Winter, Altman, and Nativ), in which segments are assigned morphosyntactic categories and are combined to form syntactic structures. Then, we use existing tools — i.e., a morphological analyzer (Segal(2000)), a part-of-speech tagger (Bar-Haim(2005); Bar-Haim et al.(2005)Bar-Haim, Sima'an, and Winter), and a general purpose parser (Schmid(2000)) — in conjunction to segment and parse unseen sentences.

The Data The data set we use is taken from the MH treebank (Sima'an et al.(2001)Sima'an, Itai, Winter, Altman, and Nativ) which consists of 5001 sentences from the daily newspaper 'ha'aretz'. We employ the syntactic categories and POS tag sets developed in (Sima'an et al.(2001)Sima'an, Itai, Winter, Altman, and Nativ). We concentrate on segmentation information and ignore inflectional morphology altogether as it would lead to extreme data sparseness. The data set we use includes 3257 sentences of length greater than 1 and less than 21. The number of segments per sentence is 60% higher than the number of words per sentence.[6] We conducted 8 experiments in which the data is split into training and test sets that are disjoint, and apply cross-fold validation to obtain robust averages.

The Morphological Analyzer A morphological analyzer helps to recover the segmentation of words by identifying their morphological constituents together with the corresponding morphosyntactic categories. Various analyses may be proposed for each word. A few standalone morphological analyzers for MH have been developed using different techniques and employing different tag sets ((Yona(2004)), (Adler and Gabai(2005)), (Segal(2000)), (Bojan(2006))). In this work, we use Segal's morphological analyzer (Segal(2000)) as it was shown to be robust and achieved the best coverage so far (96%). Since the morphosyntactic categories employed by the analyzer differ from the POS tags in the treebank, we use an automatic translation of the analyzer's output to the treebank's annotation scheme.[7]

The Part-of-Speech (POS) Tagger The most comprehensive work on POS tagging for MH to date is MorphTagger (Bar-Haim(2005)). This work uses Hidden-Markov-Models (HMMs) for POS tagging of Semitic languages. One of the tasks of MorphTagger is to pick out the segmentation of words to allow for correct POS tags' assignment. Therefore, MorphTagger uses a tri-gram model that provides short-contextual information to support disambiguation, and picks out the most probable segmentation and POS tags in context.

[6] In the complete MH corpus the average number of words per sentence is 17 while the average number of morphosyntactic segments is 26.

[7] We are grateful to Roy Bar-Haim for providing us with the script which he wrote (Bar-Haim(2005)).

The Parser To keep our preliminary exploration formally and computationally simple, we start out with a general purpose PCFG parser to which simple Maximum Likelihood (ML) estimation methods can be applied. LoPar (Schmid(2000)) is a general purpose parser for PCFGs which can be used for statistical viterbi-like parsing with any grammar or tag set. Therefore, we can use it in conjunction with the segment-based treebank grammar we obtained to parse sequences of morphological segments. Further, LoPar can parse both tagged and untagged sequences, which allows us to explore different architectural settings.

The Models We devise and implement two baseline models that are inspired by the formal account we developed in section 3.

In the first model, henceforth *Model I*, we use the morphological analyzer and MorphTagger to find the most probable segmentation for a given sentence. This is done by providing MorphTagger with multiple morphological analyses per word and letting it find the segmentation that maximizes the sum $\sum_{t_1^n} P(t_1^n, s_1^n|w_1^m)$ (Bar-Haim(2005), section 8.2). Then, the parser is used to find the most probable parse tree for the selected sequence of morphological segments. Formally, this model is an approximation of equation (1.8) (albeit a crude one, as we perform a step-wise maximization rather than making a joint decision).[8]

In *Model II* we percolate the morphological ambiguity further, to the lowest level of non-terminals in the syntactic parse trees (i.e., the POS tags). Here we use the morphological analyzer and MorphTagger in conjunction to find the most probable segmentation *and* POS tag assignment by maximizing the joint probability $P(t_1^n, s_1^n|w_1^m)$ (Bar-Haim(2005), section 5.2). Then, the parser is used to find the most probable parse tree for a sequence of segments enriched with their morphosyntactic categories. Formally, this model attempts to approximate equation (1.9). (Note that here we couple a morphological and a morphosyntactic decision, as we are looking to maximize $P(s_1^n, t_1^n|w_1^m) \approx P(t_1^n|s_1^n)P(s_1^n|w_1^m)$ (cf. equation 1.9). Then we constrain the space of possible syntactic trees to those that confine with the result of the joint maximization.)[9]

Smoothing Because of the relatively small size of our corpus (less then 10% of the WSJ portion of the Penn treebank), we encounter a sparse data problem in all levels of processing. In the current architecture, smoothing the estimated probabilities is delegated to each of the relevant subcomponents of the integrated architecture. Out of vocabulary

[8]The reason for choosing the step-wise architecture as our first model is twofold. Firstly, a step-wise architecture is computationally cheaper than a joint one, but more importantly, this is perhaps the simplest end-to-end architecture for MH parsing that one could imagine. Thus, in the lack of previous MH parsing results, it is suitable to serve as a baseline architecture against which to compare more sophisticated models.

[9]We further developed a third model, *Model III*, which is a more faithful approximation, yet computationally affordable, of equation (1.9). In *Model III* we percolate the ambiguity all the way through the integrated architecture by means of providing the parser with the n-best sequences of tagged morphological segments, and selecting the analysis $\langle \pi, t_1^n, s_1^n \rangle$ which maximizes the production $P(\pi|t_1^n, s_1^n)P(s_1^n, t_1^n|w_1^m)$. However, we have not yet obtained robust results for this model prior to the submission of this paper, and therefore we leave *Model III* for future discussion.

(OOV) words are treated by the morphological analyzer, which proposes all possible segmentations assuming that the stem is a proper noun. The Tri-gram language model used by MorphTagger is smoothed using Good-Turing discounting (the so-called 'Katz backoff', see (Bar-Haim(2005), section 6.1)),[10] and the parser uses a variant of absolute discounting, in which the discounted value is redistributed according to various backoff strategies to events with zero frequency encountered in the parsing process (Schmid(2000), section 4.4).

Evaluation We use seven measures to evaluate our integrated models. First, we present the percentage of sentences for which the model could propose a pair of corresponding morphological and syntactic analyses. This measure is referred to as *string coverage*. In order to capture tagging and parsing accuracy we refer to our redefined Parseval measures. We separate the evaluation of assigned morphosyntactic categories, i.e., *POS tags precision and recall*, and phrase-level syntactic categories, i.e., *labeled precision and recall*[11] (where root nodes are discarded as usual, and empty trees are counted as zero). Finally, we report *segmentation precision and recall*, in order to give an impression of the morphological disambiguation capabilities of the integrated model.[12]

6 Results

Table 1.2 shows the evaluation scores for the models. *Model I*, in which the parser operated on segmented sequences of words, proposed compatible morphological and syntactic analyses for 99% of the unseen sentences. However, the accuracy results are much lower – 60.3% and 58.4% labeled precision and recall for parsing, and 82.4 and 82.6% precision and recall for POS tagging.

In *model II*, the input for the parser was enriched with morphosyntactic categories that were selected in tandem with the segmentation. This improved labeled precision and recall in 0.5% and 2.1% respectively, and POS tagging precision and recall in 2.1%. However, together with the improved accuracy we observe a decrease of 3% in *string coverage*. This means that the capability of the model to provide compatible morphological and syntactic analyses has dropped. Also, we observe a decrease of 3% in our *segmentation* results, which is mainly due to the drop in *string coverage*.

[10] In this work we did not use the bootstrapping method for smoothing the lexical model nor the various heuristics for improved handling of OOV words proposed in (Bar-Haim(2005)). The reason for working with bare probabilities as estimated from the corpus is to remain faithful to the probabilities we represented in the formal exposition.

[11] Covert definite article errors are counted at the POS tags level, and discounted at the phrase-level.

[12] Since we evaluate the models' performance on an *integrated* task, sentences in which one of the subcomponents failed to propose an analysis counts as zero for *all* subtasks.

	String Coverage	Labeled Precision	Labeled Recall	POS tags Precision	POS tags Recall	Segment. Precision	Segment. Recall
Model I	99.2%	60.3%	58.4%	82.4%	82.6%	94.4%	94.7%
Model II	96.0%	60.8%	60.5%	84.5%	84.7%	91.3%	91.6%

Table 1.2: Evaluation Metrics, Models I and II

7 Analysis

This work presents a first set of statistical parsing standardized results for MH. The high string coverage score demonstrates that, in principle, models that incorporate morphological information can parse unseen sentences based on segmented and annotated corpora. Furthermore, comparison of the two models shows that coupling the morphological decision with a morphosyntactic one (currently only based on short context) improves parsing accuracy. Yet, the scores we report show that this is still insufficient for broad-coverage parsing with high accuracy comparable to other languages.

The reasons for the low parsing accuracy are several. First, the results were obtained using a relatively small set of training data, and a weak (unlexicalized) parser.[13] Further, the low accuracy is partially due to the severe ambiguity of the resulting PCFG. Since word order in MH is relatively free, CFG rules can appear in various permutations, which in turn leads to major structural ambiguity. This indicates that bare phrase structures are not adequate for capturing regularities in MH, especially with limited training data. Since we included only limited amount of morphological information that hints on possible dependencies, the parser has very limited means to recover from that.

A comparison between the models shows that while POS tags' assignment helps to improve parsing accuracy, it has negative effects on string coverage. The reason for that is that a probable yet incorrect POS tag assignment constrains the parser in a way that makes it impossible for it to recover correct syntactic structures. A POS tagger that is optimized towards syntactic decisions based on short context may result in imperfect disambiguation, especially for a language such as MH, in which long distance dependencies (e.g., due to agreement) are likely to be found.

Thus, we conclude that POS tagging is perhaps insufficient for enforcing agreement between the morphological and syntactic tasks, and propose to include larger contexts for disambiguation. We conjecture that only more extensive information sharing between the two levels of processing, i.e., morphological patterns and inflections on the one hand, and syntactic dependencies on the other hand, will allow for successful syntactic *and* morphological disambiguation.

[13] This is mainly due to the size of the corpus and its annotation scheme, which lacks head-marking.

Alphabet	aleph	bet	gimel	dalet	heh	vav	zayin	chet	tet	yod	khaf
Transliteration	a	b	g	d	h	w	z	x	j	i	k
Pronounciation	’	b,v	g	d	h	v	z	kh	t	y	k, kh
Alphabet	lamed	mem	nun	samech	’ayin	peh	tsadi	kof	reish	shin	tav
Transliteration	l	m	n	s	e	p	c	q	r	f	t
Pronounciation	l	m	n	s	’	p,ph	ts	k	r	sh, s	t

Table 1.3: Transliteration

8 Conclusion

Traditional approaches for devising parsing models and defining evaluation metrics are not adequate for MH, as they presuppose a certain language structure and separate layers of processing. Parsing Semitic languages requires serious morphological consideration, and we have shown that incorporating morphological cues (most crucially segmentation) and morphosyntactic information (currently based on short context) helps to recover parses for MH sentences. However, the high variability of the phrase structure, severe structural ambiguity, and relatively small amount of annotated data make it insufficient for completing the parsing task successfully.

Different languages mark regularities in their surface structures in different ways. English encodes regularities in word order, while MH provides useful hints for grammatical relations in its derivational and inflectional morphology. Much more work is required to prove our thesis that exploiting such information to discriminate between syntactic structures helps to correctly recover structural dependencies. In the future, we intend to develop more sophisticated models, allowing for closer interaction between morphological and syntactic processing, in order to improve parsing accuracy and facilitate morphological disambiguation.

Acknowledgments

This work is funded by the Netherlands Organization for Scientific Research (NWO), grant number 017.001.271. I would like to thank Khalil Sima’an for supervising this work, and to Remko Scha for detailed comments on an earlier draft. The Knowledge Center for Hebrew Processing at the Technion, Israel, provided corpora and processing tools, and Roy Bar-Haim provided much knowledge and technical support concerning morphological analysis and part-of-speech tagging of MH, without which this work would have been impossible.

9 Transliteration

Table 1.3 illustrates the transliteration scheme for the MH alphabet we adopt from (Sima’an et al.(2001)Sima’an, Itai, Winter, Altman, and Nativ).

Bibliography

Meni Adler and Dudi Gabai. Morphological analyzer and disambiguator for Modern Hebrew. Knowledge Center for Processing Hebrew, 2005.

Roy Bar-Haim. Part-of-speech tagging for hebrew and other semitic languages. Master's thesis, Technion, Haifa, Israel, 2005.

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

Dalia Bojan. HAMSA: A Morphological Analyzer for Hebrew Texts. The Knowledge Center for Hebrew Processing, 2006. URL `http://yeda.cs.technion.ac.il:8088/analyzer1/analyzer1.html`.

Eugene Charniak, Glenn Carroll, John Adcock, Anthony R. Cassandra, Yoshihiko Gotoh, Jeremy Katz, Michael L. Littman, and John McCann. Taggers for Parsers. *AI*, 85(1-2):45–57, 1996.

Helmut Schmid. *LoPar: Design and Implementation*. Institute for Computational Linguistics, University of Stuttgart, 2000.

Erel Segal. A Probabilistic Morphological Analyzer for Hebrew undotted texts. Master's thesis, Computer Science Department, Technion, Isreal, 2000.

Ur Shlonsky. *Clause Structure and Word Order in Hebrew and Arabic: An Assay in Comparative Semitic Syntax*. Oxford University Press, 1997.

Khalil Sima'an, Alon Itai, Yoad Winter, Alon Altman, and Noa Nativ. Building a Tree-Bank for Modern Hebrew Text. In *Traitement Automatique des Langues*, volume 42. 2001. URL `http://mila.cs.technion.ac.il/website/hebrew/resources/corpora/treebank%/index.html`.

Shlomo Yona. A Finite-state based Morphological Analyzer for Hebrew. Master's thesis, University of Haifa, Israel, November 2004.