Integrated Morphological and Syntactic Disambiguation for Modern Hebrew

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

Abstract

Current parsing models are not immediately applicable for languages that exhibit strong interaction between morphology and syntax, e.g., Modern Hebrew (MH), Arabic and other Semitic languages. This work represents a first attempt at modeling morphological-syntactic interaction in a generative probabilistic framework to allow for MH parsing. We show that morphological information selected in tandem with syntactic categories is instrumental for parsing Semitic languages. We further show that redundant morphological information helps syntactic disambiguation.

1 Introduction

Natural Language Processing is typically viewed as consisting of different layers,[1] each of which is handled separately. The structure of Semitic languages poses clear challenges to this traditional division of labor. Specifically, Semitic languages demonstrate strong interaction between morphological and syntactic processing, which limits the applicability of standard tools for, e.g., parsing.

This work focuses on MH and explores the ways morphological and syntactic processing interact. Using a morphological analyzer, a part-of-speech tagger, and a PCFG-based general-purpose parser, we segment and parse MH sentences based on a small, annotated corpus. Our integrated model shows that percolating morphological ambiguity to the lowest level of non-terminals in the syntactic parse tree improves parsing accuracy. Moreover, we show that morphological cues facilitate syntactic disambiguation. A particular contribution of this work is to demonstrate that MH statistical parsing is *feasible*. Yet, the results 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 both levels can interact and share information in both directions.

Section 2 presents linguistic data that demonstrate the strong interaction between morphology and syntax in MH, thus motivating our choice to treat both in the same framework. Section 3 surveys previous work and demonstrates again the unavoidable interaction between the two. Section 4.1 puts forward the formal setting of an integrated probabilistic language model, followed by the evaluation metrics defined for the integrated task in section 4.2. Sections 4.3 and 4.4 then describe the experimental setup and preliminary results for our baseline implementation, and section 5 discusses more sophisticated models we intend to investigate.

2 Linguistic Data

Phrases and sentences in MH, as well as Arabic and other Semitic languages, have a relatively free word order.[2] In figure 1, for example, two distinct syntactic structures express the same grammatical relations. It is typically morphological information rather than word order that provides cues for structural dependencies (e.g., agreement on gender and number in figure 1 reveals the subject-predicate dependency).

[1] E.g., phonological, morphological, syntactic, semantic and pragmatic.

[2] MH allows for both SV and VS, and in some circumstances also VSO, SOV and others.

Proceedings of the COLING/ACL 2006 Student Research Workshop, pages 49–54,
Sydney, July 2006.

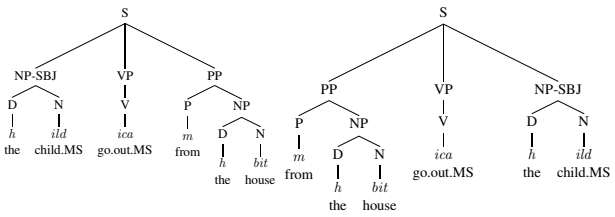

Figure 1: Word Order in MH Phrases (marking the agreement features M(asculine), S(ingular))

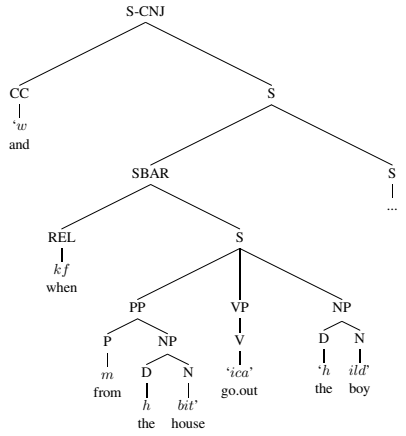

Figure 2: Syntactic Structures of MH Phrases (marking word boundaries with ' ')

Furthermore, boundaries of constituents in the syntactic structure of MH sentences need not coincide with word boundaries, as illustrated in figure 2. A MH word may coincide with a single constituent, as in 'ica'[3] (go out), it may overlap with an entire phrase, as in '$h\ ild$' (the boy), or it may span across phrases as in '$w\ kf\ m\ h\ bit$' (and when from the house). Therefore, we conclude that in order to perform syntactic analysis (parsing) of MH sentences, we must first identify the morphological constituents that form MH words.

There are (at least) three distinct morphological processes in Semitic languages that play a role in word formation. *Derivational morphology* is a non-concatenative process in which verbs, nouns, and adjectives are derived from (tri-)consonantal roots plugged into templates of consonant/vowel skeletons. The word-forms in table 1, for example, are all derived from the same root, $[i][l][d]$ (child, birth), plugged into different templates. In addition, MH has a rich array of agreement features, such as gender, number and person, expressed in the word's *inflectional morphology*. Verbs, adjectives, determiners and numerals must agree on the inflectional features with the noun they complement or modify. It can be seen in table 2 that the suffix h alters the noun 'ild' (child) as well as its modifier '$gdwl$' (big) to feminine gender. Finally, particles that are prefixed to the word may serve different syntactic functions, yet a multiplicity of them may be *concatenated* together with the stem to form a single word. The word '$wkfmhbit$' in figure 2, for instance, is formed from a conjunction w (and), a relativizer kf (when), a preposition m (from), a definite article h (the) and a noun bit (house). Identifying such particles is crucial for analyzing syntactic structures as they reveal structural dependencies such as subordinate clauses, adjuncts, and prepositional phrase attachments.

[3]We adopt the transliteration of (Sima'an et al., 2001).

a.	'*ild*'	b.	'*iild*'	c.	'*mwld*'
	$[i]e[l]e[d]$		$[i]i[l](l)e[d]$		$mw[][l](l)a[d]$
	child		deliver a child		innate

Table 1: Derivational Morphology in MH ([..] mark templates' slots for consonantal roots, (..) mark obligatory doubling of roots' consonants.)

a.	$ild\ gdwl$	b.	$ildh\ gdwlh$
	child.MS big.MS		child.FS big.FS
	a big boy		a big girl

Table 2: Inflectional Morphology in MH (marking M(asculine)/F(eminine), S(ingular)/P(lural))

At the same time, MH exhibits a large-scale ambiguity already at the word level, which means that there are multiple ways in which a word can be broken down to its constituent morphemes. This is further complicated by the fact that most vocalization marks (diacritics) are omitted in MH texts. To illustrate, table 3 lists two segmentation possibilities, four readings, and five meanings of different morphological analyses for the word-form '*fmnh*'.[4] Yet, the morphological analysis of a word-form, and in particular its morphological segmentation, cannot be disambiguated without reference to context, and various morphological features of syntactically related forms provide useful hints for morphological disambiguation. Figure 3 shows the correct analyses of the form '$fmnh$' in different syntactic contexts. Note that the correct analyses maintain agreement on gender and number between the noun and its modifier. In particular, the analysis 'that counted' (b)

[4]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).

‘*fmnh*’	‘*fmnh*’	‘*fmnh*’	‘*fmnh*’	‘*f* + *mnh*’
shmena	shamna	shimna	shimna	she + mana
fat.FS	got-fat.FS	put-oil.FS	oil-of.FS	that + counted
fat (adj)	got fat (v)	put-oil (v)	her oil (n)	that (rel) counted (v)

Table 3: Morphological Analyses of the Word-form ‘*fmnh*’

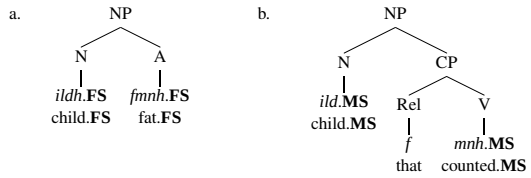

Figure 3: Ambiguity Resolution in Different Syntactic Contexts

is easily disambiguated, as it is the only one maintaining agreement with the modified noun.

In light of the above, we would want to conclude that syntactic processing must precede morphological analysis; however, this would contradict our previous conclusion. For this reason, independent morphological and syntactic analyzers for MH will not suffice. We suggest performing morphological and syntactic processing of MH utterances in a single, integrated, framework, thereby allowing shared information to support disambiguation in multiple tasks.

3 Related Work

As of yet there is no statistical parser for MH. Parsing models have been developed for different languages and state-of-the-art results have been reported for, e.g., English (Collins, 1997; Charniak, 2000). However, these models show impoverished morphological treatment, and they have not yet been successfully applied for MH parsing. (Sima'an et al., 2001) present an attempt to parse MH sentences based on a small, annotated corpus by applying a general-purpose Tree-gram model. However, their work presupposes correct morphological disambiguation prior to parsing.[5]

In order to treat morphological phenomena a few stand-alone morphological analyzers have been developed for MH.[6] Most analyzers consider words in isolation, and thus propose multiple analyses for each word. Analyzers which also attempt disambiguation require contextual information from surrounding word-forms or a shallow parser (e.g., (Adler and Gabai, 2005)).

[5]The same holds for current work on parsing Arabic.

[6]Available at `mila.cs.technion.ac.il`.

A related research agenda is the development of part-of-speech taggers for MH and other Semitic languages. Such taggers need to address the segmentation of words into morphemes to which distinct morphosyntactic categories can be assigned (cf. figure 2). It was illustrated for both MH (Bar-Haim, 2005) and Arabic (Habash and Rambow, 2005) that an integrated approach towards making morphological (segmentation) and syntactic (POS tagging) decisions within the same architecture yields excellent results. The present work follows up on insights gathered from such studies, suggesting that an integrated framework is an adequate solution for the apparent circularity in morphological and syntactic processing of MH.

4 The Integrated Model

As a first attempt to model the interaction between the morphological and the syntactic tasks, we incorporate an intermediate level of *part-of-speech (POS) tagging* into our model. The key idea is that POS tags that are assigned to morphological segments at the word level coincide with the lowest level of non-terminals in the syntactic parse trees (cf. (Charniak et al., 1996)). Thus, POS tags can be used to pass information between the different tasks yet ensuring agreement between the two.

4.1 Formal Setting

Let w_1^m be a sequence of words from a fixed vocabulary, s_1^n be a sequence of segments of words from a (different) vocabulary, t_1^n a sequence of morphosyntactic categories from a finite tag-set, and let π be a syntactic parse tree.

We define *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), where $seg(w_1^m)$ is the set of segmentations resulting from all possible morphological analyses of w_1^n.

$$s_1^{n*} = \underset{s_1^n \in seg(w_1^m)}{\text{argmax}} \; P(s_1^n | w_1^m) \qquad (1)$$

Syntactic analysis, *parsing*, identifies the structure of phrases and sentences. In MH, such tree structures combine segments of words that serve different syntactic functions. We define it formally as (2), where $yield(\pi')$ is the ordered set of leaves of a syntactic parse tree π'.

$$\pi^* = \underset{\pi \in \{\pi' : yield(\pi') = s_1^n\}}{\text{argmax}} \; P(\pi | s_1^n) \qquad (2)$$

Similarly, we define *POS tagging* as (3), where $analysis(s_1^n)$ is the set of all possible POS tag assignments for s_1^n.

$$t_1^{n*} = \underset{t_1^n \in analyses(s_1^n)}{\text{argmax}} P(t_1^n|s_1^n) \quad (3)$$

The task of the *integrated model* is to find the most probable segmentation *and* syntactic parse tree given a sentence in MH, as in (4).

$$\langle \pi, s_1^n \rangle^* = \underset{\langle \pi, s_1^n \rangle}{\text{argmax}} P(\pi, s_1^n|w_1^m) \quad (4)$$

We reinterpret (4) to distinguish the morphological and syntactic tasks, conditioning the latter on the former, yet maximizing for both.

$$\langle \pi, s_1^n \rangle^* = \underset{\langle \pi, s_1^n \rangle}{\text{argmax}} \underbrace{P(\pi|s_1^n, w_1^m)}_{\textit{parsing}} \underbrace{P(s_1^n|w_1^m)}_{\textit{segmentation}} \quad (5)$$

Agreement between the tasks is implemented by incorporating morphosyntactic categories (POS tags) that are assigned to morphological segments and constrain the possible trees, resulting in (7).

$$\langle \pi, t_1^n, s_1^n \rangle^* = \underset{\langle \pi, t_1^n, s_1^n \rangle}{\text{argmax}} P(\pi, t_1^n, s_1^n|w_1^m) \quad (6)$$

$$= \underset{\langle \pi, t_1^n, s_1^n \rangle}{\text{argmax}} \underbrace{P(\pi|t_1^n, s_1^n, w_1^m)}_{\textit{parsing}} \underbrace{P(t_1^n|s_1^n, w_1^m)}_{\textit{tagging}} \underbrace{P(s_1^n|w_1^m)}_{\textit{segmentation}} \quad (7)$$

Finally, we employ the assumption that $P(w_1^m|s_1^n) \approx 1$, since segments can only be conjoined in a certain order.[7] So, instead of (5) and (7) we end up with (8) and (9), respectively.

$$\approx \underset{\langle \pi, s_1^n \rangle}{\text{argmax}} \underbrace{P(\pi|s_1^n)}_{\textit{parsing}} \underbrace{P(s_1^n|w_1^m)}_{\textit{segmentation}} \quad (8)$$

$$\approx \underset{\langle \pi, t_1^n, s_1^n \rangle}{\text{argmax}} \underbrace{P(\pi|t_1^n, s_1^n)}_{\textit{parsing}} \underbrace{P(t_1^n|s_1^n)}_{\textit{tagging}} \underbrace{P(s_1^n|w_1^m)}_{\textit{segmentation}} \quad (9)$$

4.2 Evaluation Metrics

The intertwined nature of morphology and syntax in MH poses additional challenges to standard parsing *evaluation metrics*. First, note that we cannot use morphemes as the basic units for comparison, as the proposed segmentation need not coincide with the gold segmentation for a given sentence. Since words are complex entities that can span across phrases (see figure 2), we cannot use them for comparison either. We propose to redefine *precision* and *recall* by considering the spans of syntactic categories based on the (space-free) sequences of characters to which they correspond. Formally, we define syntactic constituents as $\langle i, A, j \rangle$ where i, j mark the location of characters. $T = \{\langle i, A, j \rangle | A \textit{ spans from } i \textit{ to } j\}$ and $G = \{\langle i, A, j \rangle | A \textit{ spans from } i \textit{ to } j\}$ represent the test/gold parses, respectively, and we calculate:[8]

$$\textit{Labeled Precision} = \#(G \cap T)/\#T \quad (10)$$

$$\textit{Labeled Recall} = \#(G \cap T)/\#G \quad (11)$$

4.3 Experimental Setup

Our departure point for the syntactic analysis of MH is that the basic units for processing are not words, but morphological segments that are concatenated together to form words. Therefore, we obtain a segment-based probabilistic grammar by training a Probabilistic Context Free Grammar (PCFG) on a segmented and annotated MH corpus (Sima'an et al., 2001). Then, we use existing tools — i.e., a morphological analyzer (Segal, 2000), a part-of-speech tagger (Bar-Haim, 2005), and a general-purpose parser (Schmid, 2000) — to find compatible morphological segmentations and syntactic analyses for unseen sentences.

The Data The data set we use is taken from the MH treebank which consists of 5001 sentences from the daily newspaper 'ha'aretz' (Sima'an et al., 2001). We employ the syntactic categories and POS tag sets developed therein. Our data set 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.[9] We conducted 8 experiments in which the data is split to training and test sets and apply cross-fold validation to obtain robust averages.

The Models *Model I* uses the morphological analyzer and the POS tagger to find the most probable segmentation for a given sentence. This is done by providing the POS tagger with multiple morphological analyses per word and maximizing 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

[7] Since concatenated particles (conjunctions et al.) appear in front of the stem, pronominal and inflectional affixes at the end of the stem, and derivational morphology inside the stem, there is typically a unique way to restore word boundaries.

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

[9] The average number of words per sentence in the complete corpus is 17 while the average number of morphological segments per sentence is 26.

probable parse tree for the selected sequence of morphological segments. Formally, this model is a first approximation of equation (8) using a stepwise maximization instead of a joint one.[10]

In *Model II* we percolate the morphological ambiguity further, to the lowest level of non-terminals in the syntactic trees. Here we use the morphological analyzer and the POS tagger 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 parse the tagged segments. Formally, this model attempts to approximate equation (9). (Note that here we couple a morphological and a syntactic decision, as we are looking to maximize $P(t_1^n, s_1^n|w_1^m) \approx P(t_1^n|s_1^n)P(s_1^n|w_1^m)$ and constrain the space of trees to those that agree with the resulting analysis.)[11]

In both models, *smoothing* the estimated probabilities is delegated to the relevant subcomponents. Out of vocabulary (OOV) words are treated by the morphological analyzer, which proposes all possible segmentations assuming that the stem is a proper noun. The Tri-gram model used for POS tagging is smoothed using Good-Turing discounting (see (Bar-Haim, 2005, section 6.1)), and the parser uses absolute discounting with various backoff strategies (Schmid, 2000, section 4.4).

The Tag-Sets To examine the usefulness of various morphological features shared with the parsing task, we alter the set of morphosyntactic categories to include more fine-grained morphological distinctions. We use three sets: *Set A* contains bare POS categories, *Set B* identifies also definite nouns marked for possession, and *Set C* adds the distinction between finite and non-finite verb forms.

Evaluation We use seven measures to evaluate our models' performance on the integrated task.

[10] At the cost of incurring indepence assumptions, a stepwise architecture is computationally cheaper than a joint one and this is perhaps the simplest end-to-end architecture for MH parsing imaginable. In the absence of previous MH parsing results, this model is suitable to serve as a baseline against which we compare more sophisticated models.

[11] We further developed a third model, *Model III*, which is a more faithful approximation, yet computationally affordable, of equation (9). There 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 it for future discussion.

	String Cover.	Labeled Prec. / Rec.	POS tags Prec. / Rec.	Segment. Prec. / Rec.
Model *I-A*	99.2%	60.3% / 58.4%	82.4% / 82.6%	94.4% / 94.7 %
Model *II-A*	95.9%	60.7% / 60.5%	84.5% / 84.8%	91.3% / 91.6%
Model *I-B*	99.2 %	60.3% / 58.4%	81.6% / 82.3%	94.2% / 95.0%
Model *II-B*	95.7%	60.7% / 60.5%	82.8% / 83.5%	90.9% / 91.7%
Model *I-C*	99.2%	60.9% / 59.2%	80.4% / 81.1%	94.2% / 95.1%
Model *II-C*	95.9%	61.7% / 61.9%	81.6% / 82.3%	91.0% / 91.9%

Table 4: Evaluation Metrics, Models *I* and *II*

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*. To indicate morphological disambiguation capabilities we report *segmentation precision and recall*. To capture tagging and parsing accuracy, we refer to our redefined Parseval measures and separate the evaluation of morphosyntactic categories, i.e., *POS tags precision and recall*, and phrase-level syntactic categories, i.e., *labeled precision and recall* (where root nodes are discarded and empty trees are counted as zero).[12] The labeled categories are evaluated against the original tag set.

4.4 Results

Table 4 shows the evaluation scores for models *I-A* to *II-C*. To the best of our knowledge, these are the first parsing results for MH assuming no manual interference for morphological disambiguation.

For all sets, parsing of tagged-segments (*Model II*) shows improvement of up to 2% over parsing bare segments' sequences (*Model I*). This indicates that morphosyntactic information selected in tandem with morphological segmentation is more informative for syntactic analysis than segmentation alone. We also observe decreasing string coverage for *Model II*, possibly since disambiguation based on short context may result in a probable, yet incorrect, POS tag assignment for which the parser cannot recover a syntactic analysis. Correct disambiguation may depend on long-distance cues, e.g., agreement, so we advocate percolating the ambiguity further up to the parser.

Comparing the performance for the different tag sets, parsing accuracy increases for models *I-B/C* and *II-B/C* while POS tagging results decrease. These results seem to contradict the common wisdom that performance on a 'complex' task de-

[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.

pends on a 'simpler', preceding one; yet, they support our thesis that morphological information orthogonal to syntactic categories facilitates syntactic analysis and improves disambiguation capacity.

5 Discussion

Devising a baseline model for morphological and syntactic processing is of great importance for the development of a broad-coverage statistical parser for MH. Here we provide a set of standardized baseline results for later comparison while consolidating the formal and architectural underpinning of an integrated model. However, our results were obtained using a relatively small set of training data and a weak (unlexicalized) parser, due to the size of the corpus and its annotated scheme.[13] Training a PCFG on our treebank resulted in a severely ambiguous grammar, mainly due to high phrase structure variability.

To compensate for the flat, ambiguous phrase-structures, in the future we intend to employ probabilistic grammars in which all levels of non-terminals are augmented with morphological information percolated up the tree. Furthermore, the MH treebank annotation scheme features a set of so-called *functional features*[14] which express grammatical relations. We propose to learn the correlation between various morphological markings and functional features, thereby constraining the space of syntactic structures to those which express meaningful predicate-argument structures.

Since our data set is relatively small,[15] introducing orthogonal morphological information to syntactic categories may result in severe data sparseness. In the current architecture, smoothing is handled separately by each of the subcomponents. Enriched grammars would allow us to exploit multiple levels of information in smoothing the estimated probabilities and to redistribute probability mass to unattested events based on their *similarity* to attested events in their integrated representation.

6 Conclusion

Traditional approaches for devising parsing models, smoothing techniques and evaluation metrics are not well suited for MH, as they presuppose separate levels of processing. Different languages mark regularities in their surface structures in different ways – English encodes regularities in word order, while MH provides useful hints about grammatical relations in its derivational and inflectional morphology. In the future we intend to develop more sophisticated models implementing closer interaction between morphology and syntax, by means of which we hope to boost parsing accuracy and improve morphological disambiguation.

Acknowledgments I would like to thank Khalil Sima'an for supervising this work, Remko Scha, Rens Bod and Jelle Zuidema for helpful comments, and Alon Itai, Yoad Winter and Shuly Wintner for discussion. The Knowledge Center for Hebrew Processing provided corpora and tools, and Roy Bar-Haim provided knowledge and technical support for which I am grateful. This work is funded by the Netherlands Organization for Scientific Research (NWO) grant 017.001.271.

[13] The lack of head marking, for instance, precludes the use of lexicalized models à la (Collins, 1997).

[14] SBJ for subject, OBJ for object, COM for complement, etc. (Sima'an et al., 2001).

[15] The size of our treebank is less than 30% of the Arabic Treebank, and less than 10% of the WSJ Penn Treebank.

References

Meni Adler and Dudi Gabai. 2005. Morphological Analyzer and Disambiguator for Modern Hebrew. Knowledge Center for Processing Hebrew.

Roy Bar-Haim. 2005. Part-of-Speech Tagging for Hebrew and Other Semitic Languages. Master's thesis, Technion, Haifa, Israel.

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

Eugene Charniak. 2000. A Maximum-Entropy-Inspired Parser. In *Proceedings of NAACL 2000*.

Michael Collins. 1997. Three Generative, Lexicalised Models for Statistical Parsing. In *Proceedings of ACL-EACL 1997*.

Nizar Habash and Owen Rambow. 2005. Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. In *Proceedings of ACL 2005*.

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

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

Khalil Sima'an, Alon Itai, Yoad Winter, Alon Altman, and Noa Nativ. 2001. Building a Tree-Bank for Modern Hebrew Text. In *Traitement Automatique des Langues*, volume 42, pages 347–380.