Accurate Unlexicalized Parsing for Modern Hebrew

Reut Tsarfaty and Khalil Sima'an

Institute for Logic, Language and Computation, University of Amsterdam
Plantage Muidergracht 24, 1018TV Amsterdam, The Netherlands

Abstract. Many state-of-the-art statistical parsers for English can be viewed as Probabilistic Context-Free Grammars (PCFGs) acquired from treebanks consisting of phrase-structure trees enriched with a variety of contextual, derivational (e.g., markovization) and lexical information. In this paper we empirically investigate the applicability and adequacy of the unlexicalized variety of such parsing models to Modern Hebrew, a Semitic language that differs in structure and characteristics from English. We show that contrary to experience with parsing the WSJ, the markovized, head-driven unlexicalized variety does not necessarily outperform plain PCFGs for Semitic languages. We demonstrate that enriching unlexicalized PCFGs with morphologically marked agreement features percolated up the parse tree (e.g., definiteness) outperforms plain PCFGs as well as a simple head-driven variation on the MH treebank. We further show that an (unlexicalized) head-driven variety enriched with the same features achieves even better performance. We conclude that morphologically rich languages introduce an additional dimension of parametrization that is orthogonal to the horizontal/vertical dimensions proposed before [11] and its contribution is essential and complementary.

Parsing Modern Hebrew (MH) as a field of study is in its infancy. Although a syntactically annotated corpus has been available for quite some time [15] we know of only two studies attempting to parse MH using supervised methods.[1] The reason state-of-the-art parsing models are not immediately applicable to MH is not only that their adaptation to the MH data and annotation scheme is not trivial, but also that they do not guarantee to yield comparable results. The MH treebank is small, the internal phrase- and clause-structures are relatively flat and variable, multiple annotated dependencies complicate the selection of a single syntactic head, and a plentiful of disambiguating morphological features are not exploited by current state-of-the-art models for parsing, e.g., English. This paper provides a theoretical overview of the MH data and an empirical evaluation of different dimensions of parameters for learning treebank grammars which break independence assumptions irrelevant for Semitic languages. We illustrate the utility of a three-dimensional parametrization space for parsing MH and obtain accuracy results that are comparable to those obtained for Modern Standard Arabic (75%) using a lexicalized parser [1] and a much larger treebank.

[1] The studies we know of are [15] which uses a DOP tree-gram model and 500 training sentences, and [16] which uses a treebank PCFG in an integrated system for morphological and syntactic disambiguation. Both achieved around 60-70% accuracy.

1 Dimensions of Unlexicalized Parsing

The factor that sets apart vanilla treebank Probabilistic Context-Free Grammars (PCFGs) [3] from unlexicalized extensions as proposed by, e.g., [10, 11], is the choice of statistical parametrization that embodies weaker independence assumptions. Recent studies on accurate unlexicalized parsing models outline two dimensions of parametrization. The first, proposed by [10], is the annotation of parent categories, effectively conditioning on aspects of a node's generation history, and the second encodes a head-outward generation process [4] in which the head is generated followed by outward Markovian sister generation processes. Klein and Manning [11] systematize the distinction between these two forms of parametrization by drawing them on a *horizontal-vertical* grid: parent-ancestor encoding is *vertical* (v) (external to the rule) whereas head-outward generation is *horizontal* (h) (internal to the rule). By varying the value of the parameters along the grid they tune their treebank grammar to achieve better performance. This two-dimensional parametrization[2] was shown to improve parsing accuracy for English [4, 1] as well as other languages, e.g., German [7] Czech [5] and Chinese [2]. However, results for languages different than English still lag behind.[3]

We claim that for various languages including the Semitic family, e.g. Modern Hebrew (MH) and Modern Standard Arabic (MSA), the horizontal and vertical dimensions of parameters are insufficient for encoding linguistic information relevant for breaking false independence assumptions. In Semitic languages, arguments may move around rather freely and the phrase-structure of clause level categories is often shallow. For such languages agreement features play a role in disambiguation at least as important as vertical and horizontal histories. Here we propose to add a third dimension of parametrization that encodes morphological features orthogonal to syntactic categories, such as those realizing syntactic agreement. These features are percolated from surface forms in a bottom-up fashion and they express information that is orthogonal to the previous two. We refer to this dimension as *depth* (d) as it can be visualized as a dimension along which parallel tree structures labeled with syntactic categories encode an increasing number of morphological features at all levels of constituency. These structures lie in a three-dimensional coordinate-system we refer to as (v, h, d).

This work focuses on MH and explores the empirical contribution of the three dimensions of parameters to analyzing different syntactic categories. We present extensive experiments that lead to improved performance as we increase the number of dimensions which are exploited across all levels of constituency. In the next section we review characterizing aspects of MH (and other Semitic languages) highlighting the special role of morphology and the kind of dependencies witnessed by morphosyntactic processes. In section 3 we describe the method and procedure for the empirical evaluation of unlexicalized parsing models for MH. In section 4 we report and analyze our results, and in section 5 we conclude.

[2] Typically accompanied with various category-splits and lexicalization.

[3] The learning curves over increasing training data (e.g., for German [7]) show that treebank size cannot be the sole factor to account for the inferior performance.

2 Dimensions of Modern Hebrew Grammar

2.1 Modern Hebrew Structure

Phrases and sentences in MH, as well as Arabic and other Semitic languages, have a relatively flexible phrase structure. Subjects, verbs and objects can be inverted and prepositional phrases, adjuncts and verbal modifiers can move around rather freely. The factors that affect word-order in the language are not necessarily syntactic and have to do with rhetorical and pragmatic factors as well. To illustrate, figure 1 shows two syntactic structures that express the same grammatical relations yet vary in their order of constituents. The level of freedom in the order of internal constituents also varies between categories, and figure 1 further illustrates that within noun-phrase categories determiners always precede nouns.[4]

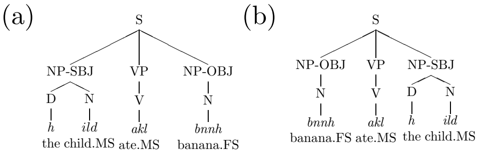

Fig. 1. Word Order in MH Phrases (marking the agreement features M(asculine), F(minine), S(ingular))

Within the flexible phrase structure it is typically morphological information that provides cues for the grammatical relations between surface forms. In figure 1, for example, it is agreement on gender and number that reveals the subject-predicate dependency. Agreement features also help to reveal the relations between higher levels of constituents, as shown in figure 2. Figure 2(a) further shows that selecting the child constituents that contribute the agreement features is not a trivial matter. Consider, for instance, definiteness in MH, which is morphologically marked (as a prefix to the stem) and behaves as a syntactic property [6]. Definite nouns exhibit agreement with other modifying phrases as shown in figure 3. The agreement on definiteness helps to determine the level of

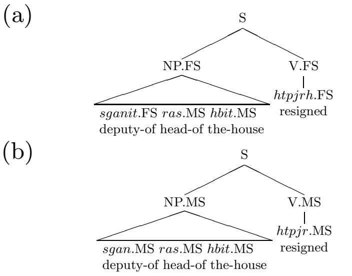
(a)
S
NP.FS
V.FS
sganit.FS *ras*.MS *hbit*.MS
deputy-of head-of the-house
htpjrh.FS
resigned
(b)
S
NP.MS
V.MS
sgan.MS *ras*.MS *hbit*.MS
deputy-of head-of the-house
htpjr.MS
resigned

Fig. 2. Phrase-Level Agreement Features (marking M(asculine), F(eminine), S(ingular))

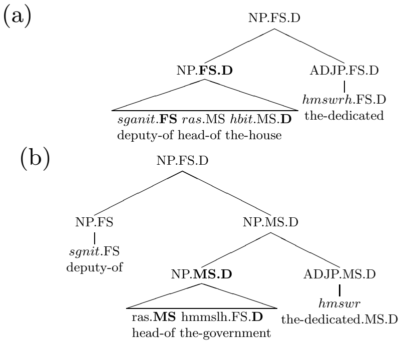

Fig. 3. Definiteness as Phrase-Level Agreement (marking M(asculine), F(eminine), S(ingular), D(efiniteness))

[4] See [17] and [8] for formal and statistical accounts of noun phrases in MH.

attachment in, e.g., the complex structure of an NP construct-state (*smixut*) or attachment to predicates in verbless sentences.[5] Figure 3(a) further illustrates that definiteness may be percolated from a different form (*hbit*.MS.**D** the-house) than the one determining the gender of the phrase (*sganit*.**FS** deputy-of).

Agreement features are thus helpful in disambiguating syntactic structures and they operate not only at the lexical level but also manifest relations between higher levels of constituents. For MH, features percolated from multiple surface forms manifest multiple kinds of dependencies and jointly determine the features of higher level constituents. Determining such features requires bi-dimensional percolation which does not coincide with head or parent dependencies, and we propose to view it as taking place along an orthogonal dimension we call *depth*.

2.2 The Modern Hebrew Treebank Scheme

The annotation scheme of the MH treebank[6] aims to capture the morphological and syntactic properties of MH we described, and differs from, e.g., the WSJ Penn treebank annotation scheme [12]. The MH Treebank is built over word segments, and the yields of the syntactic trees do not correspond to space delimited words but rather to morphological segments that carry distinct syntactic roles (i.e., each corresponding to a single POS tag). The POS categories assigned to segmented words are decorated with features such as gender, number, person and tense, and these features are percolated higher up the tree according to predefined syntactic dependencies [13]. Since agreement features of non-terminal constituents may be contributed by multiple children, the annotation scheme defines multiple dependency labels that guide the percolation of different features higher up the tree. Definiteness in the MH treebank is treated as a segment at the POS-tag level and as a feature at the level of non-terminals. As any other feature, it is percolated up the tree according to marked dependency labels. Table 1 lists the features and feature-values annotated on top of syntactic categories in the MH treebank, and table 2 describes syntactic dependencies which define the features that are to be percolated from marked child constituents.

Feature	**Value**	**Value Encoded**
gender	Z	masculine
gender	N	feminine
number	Y	singular
number	R	plural
definiteness	H	definite
definiteness	U	underspecified

Table 1. Morphological Features in the MH Treebank Annotation Scheme

Dependency Type	**Features Percolated**
DEP_HEAD	all
DEP_MAJOR	gender
DEP_NUMBER	number
DEP_DEFINITE	definiteness
DEP_ACCUSATIVE	case
DEP_MULTIPLE	all (e.g., conjunction)

Table 2. Dependency Labels in the MH Treebank Annotation Scheme

In order to comply with the flexible phrase structure in MH, clausal categories (S, SBAR and FRAG and their corresponding interrogatives SQ, SQBAR

[5] Present tense predicative sentences in MH lack a copular element.

[6] Version 2.0 of the MH treebank was made available to us in January 2007 and is currently publicly available at `http://mila.cs.technion.ac.il/english/index.html` along with a complete annotation guide and illustrative examples.

and FRAGQ) are annotated as flat structures. Verbs (VB tags) always attach to a VP mother (however only non-finite VBs can accept complements under the same VP parent). NP and PP are annotated as nested structures capturing the recursive structure of construct-state nouns, numerical expressions and possession and an additional category PREDP is added to account for sentences in MH that lack a copular element. The scheme also features null elements that mark traces and functional elements that mark, e.g. SBJ, OBJ, which we strip off and ignore throughout this study.

2.3 Treebank Grammars for Modern Hebrew

In MH there are various aspects that provide indication for the expansion possibilities of a node. Firstly, the variability in the order and number of an expansion of a non-terminal node depends on its label (e.g., while NP structures may involve nested recursive derivations, S level constituents are usually flat). Additional indication comes from the node's syntactic context. S nodes appearing under SBAR, for instance, are less shallow than those under TOP as they often involve non-finite VPs under which more modifiers can be attached. Further, although the generation of child nodes in a phrase-structure revolves, as in English, around a syntactic head, the order in which they are generated may not be as strict. Finally, morphological features indicating agreement between surface forms percolate up the tree indicating multiple dependencies. We propose to take such complementary information into account. The practice of having morphological features orthogonal to a constituency structure is familiar from theories of syntax (e.g., HPSG, LFG), however here we propose to frame it as an additional dimension for statistical estimation, a proposal which, to the best of our knowledge, has not been empirically explored before.

3 Experimental Setup

In this work we set out to empirically investigate a refined space of parameters for learning treebank grammars for MH. The models we implement use the *vertical* (v, parental history), *horizontal* (h, markovized child generation) and *depth* (d, orthogonal morphology) dimensions, and we instantiate d with the definitensess feature as it has the least amount of overlap with features determining the head.

We use version 2.0 of the MH treebank [15] which consists of 6501 sentences from the daily newspaper 'ha'aretz' and employ the syntactic categories, POS categories and morphological features annotated therein. The data set is split into 13 sections consisting of 500 sentences each. We use the first section (section 0) as our development set and the last section (section 12) as our test set. The remaining sentences (sections 1–11) are used for training. After cleaning the data set we remain with a *devset* of 483 sentences (average length in word segments 48), a *trainset* of 5241 sentences (53) and a *testset* of 496 sentences (58).[7]

[7] Since this work is only the first step towards the development of a broad-coverage statistical parser for MH (and other Semitic languages) we use only the development set and leave our test set untouched.

Lexicalize	select and percolate lexical heads and their categories for markovization
Linearize	linearize RHS of CFG productions (using [9])
Decorate	annotate contextual/morphological features on top of syntactic categories

Table 3. Transforms over the MH Treebank

Name	Params	Description	Transforms used
DIST	$h = 0$	0-order Markov process	**Lexicalize(category), Linearize(distance)**
MRK	$h = 1$	1-order Markov process	**Lexicalize(category), Linearize(distance, neighbor)**
PA	$v = 1$	Parent Annotation	**Decorate(parent)**
DEF	$d = 1$	Definiteness Percolation	**Decorate(definiteness)**

Table 4. Implementing Different Parametrization Options using Transforms

Our methodology is similar to the one used by [10] and [11]. We transform our training set according to certain parametrization decisions and learn different treebank grammars according to different instantiations of one, two, and three dimensions of parameters (tables 3 and 4 show the transforms we use to instantiate different parameters).

The input to our parser is a sequence of word segments (each corresponding to a single POS-tag). This setup assumes partial morphological disambiguation (e.g., segmentation) but we do *not* provide the parser with POS tags information.[8] We train a treebank PCFG on the resulting treebank using relative frequency estimates, and we use BitPar, an efficient general-purpose PCFG parser [14], to parse unseen sentences.[9]

We evaluate our models using EVALB focusing on bare syntactic categories. We report the average F-measure for sentences of length up to 40 and for all sentences ($F_{\leq 40}$ and F_{All} respectively), once including punctuation marks (WP) and once excluding them (WOP). For selected models we show a break-down of the average F_{All} (WOP) measure for different categories.

4 Results and Analysis

In a series of experiments we evaluated models that instantiate one, two or three dimensions in a coordinate-system defined by the parameters (v, h, d). We set our baseline model at the $(0, 0, 0)$ point of the coordinate-system and compared its performance to a simple treebank PCFG and to different combinations of parameters. Table 5 shows the accuracy results for parsing section 0 for all models. The first outcome of our experiments is that our head-driven baseline performs slightly better than a vanilla treebank PCFG. Because of the variable phrase-structure a simple PCFG does not capture relevant generalization about sentences' structure in the language. However, enriching a vanilla PCFG with orthogonal morphological information (definiteness in our case) already performs better than our baseline unlexicalized model. In comparing the contribution of three one-dimensional models we observe that the depth dimension contributes

[8] This setup makes our results comparable to parallel studies in other languages.

[9] We smooth pre-terminal rules by providing the parser with statistics on "rare words" distribution. The frequency defining "rare words" is tuned empirically and set to 1.

the most to parsing accuracy. These results demonstrate that incorporating dependency information marked by morphology is important to analyzing syntactic structures at least in as much as the main head-dependency is. The results for two-dimensional models re-iterate this conclusion by demonstrating that selecting the depth dimension is better than not doing so. Notably, the configuration most commonly used by current state-of-the-art parsers for English (i.e., $(v, h, 0)$, cf. [11]) performs slightly worse than the ones incorporating a depth feature. A three-dimensional annotation strategy achieves the best accuracy results among all models we tested.[10] The error reduction rate from a plain PCFG is more than 20%, providing us with a new, much stronger, lower bound on the performance of unlexicalized treebank grammars in parsing MH.

The general trend observed in our results is that higher dimensionality is better. Different dimensions provide different sorts of information which are complementary. As further illustrated in table 6 the internal structure of different syntactic constituents may benefit to a different extent from information provided by different dimensions. Table 6 shows the breakdown of the F_{All}(WOP) accuracy results for the main syntactic categories in our treebank. In the lack of parental context ($v = 0$) the Markovian head-outward process ($h = 1$) encodes information relevant for disambiguating the flat variable phrase-structures. The morphological dimension ($d = 1$) helps to determine the correct labels and attachment via the agreement with modifiers within NP structures. In the presence of a vertical history ($v = 1$) that provides cues for the expansion possibilities of nodes, the contribution of an orthogonal morphological feature ($d = 1$) is even more substantial. Accuracy results for phrase-level categories (ADJP, ADVP NP and VP) are better for the v/d combination than for the v/h one. Yet, high-level clausal categories (S and SBAR) benefit from head-outward markovization processes ($h = 1$) which encode additional rhetoric, pragmatic, and perhaps extra linguistic knowledge that govern order-preferences in the genre.

Name	Params (v, h, d)	F_{ALL} WP	$F_{\leq 40}$ WP	F_{ALL} WOP	$F_{\leq 40}$ WOP
BASE	$(0, 0, 0)$	66.56	68.20	67.59	69.24
PCFG	$(0, \infty, 0)$	65.17	66.63	66.17	67.7
PCFG+DEF	$(0, \infty, 1)$	67.53	68.78	68.7	**70.37**
PA	$(1, 0, 0)$	68.87	70.48	69.64	70.91
MRK	$(0, 1, 0)$	66.69	68.14	67.93	69.37
DEF	$(0, 0, 1)$	68.85	69.92	70.42	**71.45**
PA+MRK	$(1, 1, 0)$	69.97	71.48	70.69	71.98
MRK+DEF	$(0, 1, 1)$	69.46	70.79	71.05	72.37
PA+DEF	$(1, 0, 1)$	71.15	72.34	71.98	**72.91**
PA+MRK+DEF	$(1, 1, 1)$	72.34	73.63	73.27	**74.41**

Table 5. Multi-Dimensional Parametrization of Treebank Grammars (Head-Driven Models are Marked $h \neq \infty$): $F_{\leq 40}, F_{ALL}$ Accuracy Results on Section 0.

(v, h, d)	$(0, 0, 1)$	$(0, 1, 0)$	$(1, 0, 1)$	$(1, 1, 0)$
	$v = 0$		$v > 0$	
ADJP	76.42	**76.62**	**81.34**	80.12
ADVP	72.65	**74.77**	**79.66**	78.19
NP	**75.28**	74.74	**79.29**	77.66
VP	71.10	**71.80**	**75.69**	73.89
S	74.41	**78.08**	76.04	**79.49**
SBAR	56.65	**63.62**	59.59	**65.65**
SQ	50.00	**54.55**	**44.44**	40.00
FRAG	**56.00**	53.85	**61.02**	58.62

Table 6. The Contribution of the *horizontal* and *depth* Dimensions ($v > 0$ Marks Parent Annotation, $h > 0$ Marks 1-Order Markov Process): F_{All} (WOP) per Syntactic Category on Section 0

[10] The addition of an orthogonal *depth* dimension to the *horizontal-vertical* space goes beyond mere "state-splits" (cf. [11]) as it does not only encode refined syntactic categories but also signals linguistically motivated co-occurrences between them.

5 Conclusion

Tuning the dimensions and values of the parameters in a treebank grammar is largely an empirical matter, but our results point out that the selection of parameters for statistical estimation should be in tune with our linguistic knowledge of the factors licensing grammatical structures in the language. Morphologically rich languages introduce an additional dimension into the expansion possibilities of a node which is orthogonal to the vertical [10] and horizontal [4] dimensions systematized by [11]. Via a theoretical and empirical consideration of syntactic structures and morphological definiteness in MH we show that a combination of multiple orthogonal dimensions of parameters is invaluable for boosting the performance of unlexicalized parsing models. Our best model provides us with a new, strong, baseline for the performance of treebank grammars for MH.

Bibliography

[1] D. Bikel. Intricacies of Collins' Parsing Model. *Computational Linguistics*, 30(4), 2004.

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

[3] E. Charniak. Tree-Bank Grammars. In *AAAI/IAAI, Vol. 2*, pages 1031–1036, 1996.

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

[5] M. Collins, J. Hajic, L. Ramshaw, and C. Tillmann. A Statistical Parser for Czech. In *Proceedings of ACL*, College Park, Maryland., 1999.

[6] G. Danon. Syntactic Definiteness in the Grammar of Modern Hebrew. *Linguistics*, 39(6):1071–1116, 2001.

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

[8] Y. Goldberg, M. Adler, and M. Elhadad. Noun Phrase Chunking in Hebrew: Influence of Lexical and Morphological Features. In *Proceedings of COLING-ACL*, 2006.

[9] F. Hageloh. Parsing using Transforms over Treebanks. Master's thesis, University of Amsterdam, 2007.

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

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

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

[13] A. Milea. Treebank Annotation Guide. MILA, Knowledge Center for Hebrew Processing, 2007.

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

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

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

[17] S. Wintner. Definiteness in the Hebrew Noun Phrase. *Journal of Linguistics*, 36:319–363, 2000.